Joint Modelling Of Longitudinal And Time-to-event Data

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by:

Matthew Alan Powney

October 2015

Joint Modelling of Longitudinal and Time-to-event data - Matthew Powney

Supervised by Ruwanthi Kolamunnage-Dona and Paula Williamson

Abstract

A randomised control trial (RCT) is considered to be the gold standard for investigating the efficacy of new and novel treatments. However, in RCTs with longitudinal outcomes and high percentages of dropout, a poor handing of missing data can be problematic when trying to establishing the efficacy of an intervention. Joint modelling of longitudinal and time-to-event data is a novel methodology that can be used to monitor a longitudinal outcome while simultaneously accounting for time-to-dropout. This is achieved using a mean zero latent Gaussian process, and relies on the estimation of the parameter γ which models the association between the longitudinal and time-to-event components. However, joint modelling is still a relatively new topic for research. The aim of this thesis is to provide and develop a greater understanding for both the design and analysis elements of joint modelling.

In Chapter 2 a simulation study to test the success of various missing data handling methods is presented. This demonstrated that for RCTs with missing data, joint modelling performs as well as the common alternative methods when estimating longitudinal treatment effect. Despite these benefits, a systematic review conducted in Chapter 3 showed that Joint Modelling is rarely used in practice in RCTs with longitudinal outcome data. One contributing factor to the underuse of joint models may be the lack of understanding and research into sample size calculations for a trial using joint modelling. In Chapter 4, sample size formulae are derived for β_2 and γ in the Henderson et al. (2000) random slope and intercept specification of the joint model. These sample size and power calculations depend on knowledge about the value of γ in a trial. Currently, the understanding of the interpretation of γ is limited, and no previous investigations into the relationship between magnitude of γ and change in longitudinal outcome for dropouts has been carried in published literature. In Chapter 5, a visualisation of this relationship is presented.

In Chapter 6, software is developed in R to carry out and apply some diagnostic procedures for joint models. Chapter 7 demonstrates the applicability of the methods described in this thesis, in which joint modelling is utilised to analyse a wide range of datasets.

Contents

Abbreviations

ASS	Asthma Severity Score
CONSORT	Consolidated Standards Of Reporting Trials
EM	Expectation-Maximisation
LOCF	Last Observation Carried Forwards
MAR	Missing At Random
MCAR	Missing Completely At Random
MICE	Multiple Imputation by Chained Equations
MNAR	Missing Not At Random
MSE	Mean Square Error
RCT	Randomised Control Trial
VIF	Variance Inflation Factor
WLSM	Weighted Least Squares Method

Chapter 1

Introduction

1.1 Introduction

A randomised control trial (RCT) is considered to be the gold standard for investigating the efficacy of new and novel medical interventions [1]. However, designing and running an RCT is an intricate and lengthy procedure, as these studies are prone to bias. Even in well-designed trials, some problems may inevitably arise that need to be addressed in order to ensure that the results of a trial are applicable in a wider context. One such issue is the occurrence of missing data [2], as it is the ethical prerogative of a patient randomised within a trial to withdraw or dropout from an RCT at their discretion [3]. However, if a large number of patients dropout from a trial this can prove problematic when establishing the efficacy of an intervention.

In many trials with missing data present, the standard approach is to only include patients that completed the measurement schedule, which can lead to biased results. In more recent times an appreciation has developed for the need to adjust the standard statistical analyses to compensate for missing data [4]. The topic of methods for missing data handling has become widely discussed in the biostatistical community, where both pre-analysis methods i.e. imputation [5–7] and models which compensate for missingness i.e. mixed models, joint models [8, 9] have been developed. Designing and running an RCT is both a time consuming and financially costly procedure, and an inaccurate assessment of treatment effect due to mishandling or ignoring missing data can lead to erroneous conclusions being drawn [4].

In this chapter, the basic foundations for designing and running a clinical trial are presented and the potential causes of bias are discussed. The main topic of the thesis, joint modelling of longitudinal and time-to-event data, is introduced alongside the MAGNETIC trial, which acts as a motivating dataset for the work. Joint modelling of longitudinal and time-to-event data is a flexible novel methodology that can be used to model a longitudinal

outcome over time alongside the time to a clinical event or dropout. [10].

1.2 Randomised Control Trials

In the past 100 years, great developments have been made in the area of medicine. An increase in knowledge and resources has led to the current advanced climate. The rise of scientific methods along with the developments in biological and medical research have saved an immeasurable number of lives, and improved prognoses for billions of people. The development and refinement of the RCT has played a large part in these medical advances [11]. An RCT is a prospective experiment in a clinical environment in which interventions are tested on human subjects for the benefit of evaluation [12]. Each individual in a trial is prescribed a treatment based upon a random mechanism, the effect of the intervention on the prognosis of the patients is monitored and the results from each treatment group are compared at the end of the study using a pre-specified statistical analysis plan [13].

Conclusive results from an RCT are more reliable than any other type of clinical research study according to a classification provided by Sackett et al, 2000 [14]. However, inappropriate or lacklustre RCT designs and analyses can lead to incorrect results [15]. The main objective of a trial is to obtain clinical results from a sample of patients which are reflective of the treatment effects when applied to a general population [16] and, in order to accomplish this, every aspect of an RCT needs to be carefully controlled to ensure the results obtained are reflective of a wider audience. To aid with these issues, guidelines have been published on the design and analysis of RCTs [17]. In Section 1.2.1 the general format of an RCT is discussed as well as potential problems that could arise.

1.2.1 General Randomised Control Trial Characteristics and Considerations

The first recorded RCT was conducted in 1931 to evaluate sanocrysin for pulmonary tuberculosis [18]. In the trial, 24 patients were randomised to 2 groups using the flip of a coin and the outcomes for each patient was analysed. While the basic premise is the same, in modern day trials a greater attempt is made to avoid bias and ensure the safety of participants which makes the designing of RCTs a more detailed procedure.

The intention of an RCT is to compare the results from a novel treatment against a control within a sample of patients that reflect a whole population and thereby improve understanding of a disease and its prevention or treatment [15]. The first step is to identify a potentially beneficial intervention to a given ailment. After preliminary testing has been done on the chemical and medical properties of the intervention, some are eligible for greater investigation in an RCT setting. To ensure that patients are not exposed to unnecessary risks during these studies, trials are carried out in 4 different phases [19–21].

- **Phase I** - A preliminary study designed to investigate the toxicity and safety of an intervention. These are carried out on a small number of patients with a view to establishing the optimal dose and to test for potentially harmful side effects.
- **Phase II** - Further tests are carried out on a larger number of patients who are ill to assess whether the benefits of the intervention outweigh the risks.
- **Phase III** - Interventions that progress through phases I and II are tested and compared against the standard treatment (control) or a placebo in a larger scale trial. This is commonly to establish whether the new novel treatment outperforms or is as good as the current treatment being used. In a two-arm trial patients are randomly allocated to one of the two treatments. These trials typically take more than a year to conduct.

- **Phase IV** - After a treatment has progressed through phases I - III and been granted marketing licences, phase IV trials are carried out. These may be to establish late effects or morbidity when patients have been given a treatment, as well as monitoring previously identified side effects and effectiveness on a larger scale.

By approaching trials in this way, harmful and non-beneficial treatments can be established within the smaller scale phase I and phase II studies before they are exposed to a wider audience.

To ensure the transparency of a trial and a reduction in the risk of bias a trial protocol is developed and prepared prior to the commencement, which outlines various aspects of the trial design. These include a clear definition of outcome, eligibility criteria, sample size, methods of randomising patients to each treatment group and blinding [1].

Outcomes

The outcome recorded in a clinical trial should be well defined, with a clear interpretation in a clinical environment. If it is difficult to interpret or ambiguous, then when statistical analyses are carried out the prognostic value of the outcome may be redundant. Similarly it should apply to the research questions posed in the trial protocol [1] and be both attainable and unbiased [19]. In the cases where an outcome is difficult to measure in some patients, this would both require a larger sample size in a study and could potentially exclude a proportion of the population which could influence the results. Therefore, the methods of outcome measurement should be clearly stated in the RCT protocol [19]. These considerations apply to all trial designs and properties. However, additional trial specific considerations should be addressed dependent on the nature of the study.

In this thesis, the majority of the work will be related to trials with longitudinal data. Longitudinal data is defined as outcomes that are measured repeatedly over time [22], and typically these outcomes are categorical, discrete or continuous. For these type of trials,

extra considerations have to be made at the design stage. By developing a measurement schedule that may suit a large proportion of patients, missing data can be minimised [23]. The main focus for this thesis will be longitudinal studies with a balanced design, in which longitudinal responses are recorded at the same times for each patient [24].

In some trials, the primary interest is not in a measurement recorded at a single or multiple timepoints after randomisation, but the time taken for an event to occur. This is known as time-to-event data. Some frequently used time-to-event outcomes in RCTs are time-to-death, time-to-remission in cancer trials and time-to-dropout. The statistical techniques used for this type of analysis are vastly different to longitudinal outcomes or standard linear modelling analyses of treatment effect. One method used to model this type of data is the Cox-Proportional Hazards model [95], which is discussed later in the chapter.

Eligibility Criteria

In the trial protocol, the eligibility criteria for patients is also defined. This is the list of conditions that need to be satisfied for a patient to be allowed into a study [1]. It should be a sample of individuals comparable to the entire population that an intervention is aimed to treat. However, this is not the only factor to consider. The safety of patients is also important and often the eligibility criteria will exclude any patients that are at greater risk of adverse effects. For example, many trials do not include pregnant women in non pre-natal studies [25]. To ensure the ethics of an RCT are upheld, in general, patients included in a study should be capable of giving true informed consent [26].

Sample Size

The number of patients inducted into a study is determined by a multitude of factors including the nature of the condition, the trial design, the desired precision of the results and the availability of participants [15]. Before a study is carried out, a power and significance

level is set for the trial. The power is defined as the probability of a difference in treatment effect being detected, given that one exists [27], and the higher the sample size the greater the power. The approach to establishing sample size may be different depending on the aims of the study and the methods of analysis. For example an equivalence trial [28] will require a different number of patients than a superiority trial [29], or trials with a longitudinal outcome recorded over time will require different numbers of patients to a trial with an outcome recorded once. Based on all these factors, a sample size for the trial is derived using mathematical methods, and included in the trial protocol prior to the randomisation of patients. In some cases where closed form estimates of the sample size are not obtainable, simulation methods are used to establish the sample size. Greater discussions about the factors effecting sample size are discussed in Chapter 4, however these include the true difference in treatment effect, the significance level, the aforementioned power of a trial and the specific trial design.

Randomisation

Randomisation is the method by which patients admitted to a clinical trial are assigned to one of the interventions [15]. In order for the results of a trial to be valid it is important that there are no important prognostic differences between the patients in each treatment group. To ensure this, there are many different randomisation techniques to assign treatments. The following is a selection of the most common methods [30];

- **Simple Randomisation** - For example flipping a coin, as used in the first ever recorded RCT, or using random number tables to allocate patients to each treatment group.

- **Stratified Randomisation** - Stratified randomisation ensures that equal or close to equal numbers of patients with a chosen characteristic are allocated to each treatment arm. This characteristic is usually something which may have an effect on prognosis.

- **Block Randomisation** - Blocks of random sequences are generated to assign patients to different treatments. For example in a block of six, the patients in order of recruitment could be allocated to treatment group's ABBAAB. In this case the first patient admitted to a study would be placed in treatment group A, the second and third treatment group B etc. After the first 6 patients an alternative block can be generated to allocate patients 7-12, e.g BAAABB. This will ensure a similar number of patients are randomised to each group. There are many software programs that are capable of carrying out this allocation process.

- **Pairwise Randomisation** - Two patients with similar prognostic factors are coupled and one allocated to treatment group A, and the other to treatment group B. Like block randomisation this ensures a balance in the groups in terms of numbers, but also in important prognostic factors.

- **Minimisation** - Minimisation uses algorithmic procedures to ensure that the treatment groups are balanced in terms of important prognostic factors of the patients.

Clearly there are advantages and disadvantages to using each method, and the appropriateness in each case may be dependent on the specific aims and requirements of the trial. Using simple randomisation is a straight forward procedure, however it can lead to an imbalance in the number of patients per treatment group [31]. Similarly there are issues with the potentially non random nature of stratified randomisation, blocking randomisation in smaller scale clinical trials and the difficulty in finding two patients with similar prognostic factors in pairwise randomisation [31]. To ensure transparency, the eligibility of patients should be established before randomisation and baseline readings should also be taken before treatment groups are assigned [1].

Blinding

In the first clinical trial in 1931, one phenomenon which the trialists were almost certainly unaware of was the placebo effect. The placebo effect is a beneficial effect produced by a

placebo or treatment which is not accounted for by the medical properties of the intervention itself [15]. In layman's terms, patients who believe a treatment will be successful may display different prognostic results to those who believe the treatment will be unsuccessful. In RCT analysis it is important to obtain data which eliminates a disparity caused by the placebo effect across treatment groups as much as possible, which can be achieved by blinding the treatment.

The following are types of blinding in RCTs [15]:

- **Single Blind** - The patient is unaware of whether they are a member of the control or experimental group in a trial, or which intervention they are receiving. However, the physician is aware of these details.
- **Double Blind** - Both the patient and the doctor are unaware what treatment is being administered to each individual. This reduces the potential for physicians to manipulate the results and reduce the influence of the placebo effect.
- **Triple Blind** - The patient, the physician and the data analyst/statisticians are all unaware of which group was administered each treatment. This addresses not only the problems that can arise from the physician and patient having knowledge of treatments, but also reduces the potential for reporting bias.
- **Open** - Both the doctor and the patient are aware of the treatment being received. In some instances blinding in a trial may not be possible. For example it can be difficult to blind a trial where a surgical treatment is being compared to a pill.

Statistical Analysis Plan

After the methods of blinding have been chosen, sample size formulae calculated and patients have been randomised to an intervention, the follow up of patients can commence.

Relevant outcomes are collected and the patients are monitored for the entire duration of the trials.

The nature of the statistical analysis should be discussed in detail in the trial protocol and a statistical analysis plan should be approved prior to conducting the analysis. In general, RCTs can be separated into one of three types; superiority, equivalence and non-inferiority trials. Each of these will require different guidelines for statistical analysis. Superiority trials are the most common, in which the aim of a trial is to prove a new novel treatment is better than the current gold standard treatment by observing an outcome [29]. In superiority trials we generally look for a treatment effect which is significantly better than the control at the 95% significance level. In equivalence trials the aim is to show that the novel treatment is "equivalent" to the current standard treatment [28]. For this we define a two sided equivalence margin, which is centred around the treatment effect estimate of the control. If the confidence interval for the treatment effect of the novel treatment falls within this margin, the treatments can be deemed clinically equivalent. This term is defined clinically in the trial protocol. By showing two treatments are clinically equivalent, it may be possible to establish that the new novel treatment is more appropriate for the population if the side effects for the new treatment are less severe. Likewise the aim of a non-inferiority trial is to show that the novel treatment is "no worse" than the standard treatment used [32]. For this we establish a one sided inferiority margin, and if the confidence interval for the new treatment effect is shows the novel treatment to be better than the effect of the control minus this inferiority margin at the 95% level, a novel intervention is said to be non inferior. The majority of the research in this thesis will be applied to superiority trials.

For trials with repeated measures data, there are many longitudinal data analysis approaches that can be employed. Arguably the most common of those utilised in clinical trial literature are linear mixed models [33], which are presented in greater detail in Chap-

ter 2, Section 3.3. These linear mixed models are capable of taking into account both fixed effects (constant across all patients) and random effects (unique to each individual patient) in order to estimate the parameter of treatment effect, while accounting for individual patient variation. One advantage of carrying out a longitudinal trial is that a greater amount of clinical information can be gained by observing the individual outcome trajectories of patients [22]. Consequently a greater understanding of the effect of the treatment can be obtained. After results have been collected for an RCT, conclusions are drawn as to the clinical significance of a treatment, based upon the parameter estimates and statistical significance.

1.2.2 Sources of Bias

Section 1.2.1 provides a description of how RCT protocols and procedures are used to maximise the applicability of results within a trial. However, despite the rigorous guidelines, it is impossible to ensure a trial is completely free from bias. Trialists need to be mindful of potential sources of bias so that these can be addressed in the trial design stage. The following is a summary of the most common types of bias in RCTs. [34]

- **Selection Bias** - This arises when there are characteristic prognostic differences between the patients in each treatment group, which can be caused by use of incorrect randomisation methods. This imbalance can result in a difference in performance for the interventions and provides an inaccurate reflection of the comparative treatment effects when applied to the general population.

- **Performance Bias** - This is caused by a difference in the care provided to patients in each treatment group, which may be due to insufficient blinding methods.

- **Detection Bias** - When outcomes are determined using different methods of measurement in different treatment groups there may be discrepancies within the results.

- **Attrition Bias** - This is caused by a misrepresentation of the data which is induced by inappropriate methods of dealing with the missingness in a study. For example,

if patients who dropped out from the study in one treatment arm are prognostically different to those in another, ignoring all patients that dropped out can provide skewed results in an analysis.

- **Reporting Bias** - This arises from selective reporting of statistical results and manipulation of results to fit a given conclusion. A subset of this, "within-study publication bias"', also known as outcome reporting bias, has become one of the main target areas in clinical trial research as it has been identified as one of the most substantial biases affecting results from individual studies [35].

- **Other Biases** - Along with the common biases, some biases may be specific to certain outcomes or specific trials. These non generalisable biases may be dealt with on a case by case basis.

One of the main areas of focus for this thesis is dealing with the problem of attrition bias caused by a mishandling of missing data. Methods to address these problems are discussed in detail in Chapter 2.

1.3 MAGNETIC trial

MAGNETIC is a recently conducted double blind Phase III RCT with longitudinal measurements taken, which tested the efficacy of adding magnesium nebuliser treatment to the standard asthma treatment for children with acute severe asthma. Many of the studies and investigations in this thesis will be based around and motivated by the circumstances and data from this trial.

In 2011, figures show that there were 1,167 deaths as a result of asthma in the UK and one of the main reasons for acute hospital admission was acute asthma in children. Studies indicate that one of the main risk factors leading to death for asthma is inadequate treatment of the disease, and up to 90% of deaths are considered to be preventable [36].

The standard treatment for acute severe asthma in children is a nebuliser containing a combination of salbutamol and ipratropium bromide. The dose is dependent on the age of the child in question [37]. Early phase studies had showed that the inclusion of magnesium sulfate in the nebuliser may have a clinically positive effect on the severity of asthma score. The aim of the MAGNETIC trial was to test this hypothesis in children with acute severe asthma.

The outcome of interest in the MAGNETIC trial was the disease severity, which was measured by Yung asthma severity score (ASS) [38]. ASS is measured on a scale ranging from 0 to 9 and contains three equally weighted components; wheeze score, muscle use and heart rate. Each patient was given a rating of 0 to 3 for each of the subcomponents and the sum of these is defined as the asthma severity score. A higher ASS indicates more severe symptoms.

In MAGNETIC, ASS was measured repeatedly over time. The intervention was administered to each patient for the first 60 minutes post randomisation, with patients in the control group being given standard treatment + placebo and the experimental group given standard + magnesium sulphate. ASS was measured at baseline, 20 minutes, 40 minutes and 60 minutes during treatment, and at follow up times of 120 minutes, 180 minutes and 240 minutes. The primary outcome was ASS score after 60 minutes, however in this thesis the focus will be on monitoring the ASS score over time in a longitudinal context.

Children from the age of 2-15 that showed signs of acute severe asthma were admitted to the study from 30 hospitals in the UK. To reduce the chance of bias, block randomisation was used and stratified by hospital and the trial was triple blinded so that the patient, physician and data analyst were unaware of which treatment group corresponded to which intervention. Over a period of 28 months, 508 children were admitted into the trial with 256 being given the standard treatment + placebo and 252 were randomised to standard

treatment + $McSO_4$.

Patients were followed through until 240 minutes and after the trial, a data analysis was carried out using t-tests to establish the difference between treatment groups after 60 minutes. This analysis showed that in the treatment group with magnesium added, the ASS was approximately 0.25 (2.77 s.d.) lower on average after 60 minutes, which was statistically significant. However, the MAGNETIC trial also had a number of patients that withdrew or dropped out of the study, which could potentially skew the results. The following table represents the amount of missing values at each time point:

	Time (minutes)					
Group	20	40	60	120	180	240
Magnesium	7 (2.7%)	13 (5.1%)	24 (9.4%)	27 (10.5%)	44 (17.2%)	60 (23.4%)
Placebo	3 (1.2%)	8 (3.2%)	12 (4.8%)	18 (7.1%)	27 (10.7%)	35 (13.9%)

Table 1.1: Number of missing values of patients in MAGNETIC

By the end of follow up, there was found to be 26.7% of patients that had missing data at at least one time point, which we will define in this thesis as a dropout (32.6% in the magnesium group and 20.3% in the placebo group). The reasons for dropout included children being discharged due to positive prognosis, children being discharged due to a negative effect of the treatment and causes unrelated to prognosis (for example children not liking the taste of the nebuliser). To have a greater understanding of the prognostic effect of adding magnesium nebuliser to the standard acute severe asthma treatment, a longitudinal data analysis of this trial would be beneficial. However, due to the diversity of the reasons for dropout, excluding missing data from the analysis may provide misleading conclusions. From a clinical perspective it is of interest to be able to model the change in longitudinal outcome over time while accounting for patients that dropped out.

1.4 Joint Modelling of Longitudinal and Time-to-event Data

One method of accounting for these types of missingness within the data is joint modelling of longitudinal and time-to-event data [39]. In many studies, the time at which a patient experiences a pre-defined clinical event, or the time-to-dropout is recorded alongside some longitudinal measurements. Joint modelling is a novel methodology which takes into account patient events or dropout while monitoring a longitudinal outcome over time.

1.4.1 History

The concept of joint models stem from a paper written in 1995 by Tsiatis et al. which was motivated by the difficulty in establishing a reliable surrogate marker for disease progression in AIDS [40]. The aim of this paper was to establish whether CD4 count could be used as a surrogate marker for AIDS in HIV trials to establish the efficacy of a treatment more efficiently. At the time, a clear correlation between CD4 count and disease progression had been established in the literature [41], but a greater understanding of the outcome needed to be provided before this knowledge could be applied in a clinical environment. Large biological variation and CD4 counts only being measured on rare occasions rendered standard models invalid. This led to the innovative approach of developing a two stage joint model. This two stage approach allowed longitudinal biomarkers to be taken into account when modelling for a survival outcome. The first stage involved modelling the counts using a repeated measures random components model. A derivation was then made in the second stage for estimating the parameters of a Cox Model to establish a relationship for the time-to-event. The results of the paper showed that CD4 count was an appropriate surrogate marker for AIDS, which has proved to be valuable in this area of research.

In 1996 this concept was developed further by Faucett and Thomas, who proposed a random effects model with proportional hazards to monitor the event time to jointly es-

timate data of this nature [43, 44]. Wulfsohn and Tsiatis developed this approach further by considering the scenario where data was measured infrequently [45]. In this paper, the expectation maximisation (EM) algorithm was used to estimate parameters by maximising the joint likelihood of a Cox-Proportional Hazards model with longitudinal covariates based upon random effects models incorporated [42]. Henderson, Diggle, and Dobson (2000) built upon this work by introducing a stationary Gaussian process, which also utilises the EM algorithm to estimate longitudinal and event-time components, alongside an association parameter γ. Software is available to fit models of this type in R, using the joineR package [46]. Greater details of this model specification are presented in Section 1.4.3 and Section 2.3.4. Further joint model classifications have been made, by Ibrahim, Chen, and Sinha (2004) [49], Xu and Zeger (2001b) [51], and Song, Davidian, and Tsiatis (2002) [52] who all extended the longitudinal model to the multivariate case using both Bayesian and Frequentist methods.

1.4.2 Aims of Joint Modelling

There are three primary aims and inferences that can be drawn from joint models. These are:

- Inferences about longitudinal measurements being adjusted to compensate for informative events or dropout.
- Modelling of time-to-event or dropout while compensating for longitudinal measurements.
- Analysis of how the model evolves with respect to both the longitudinal measurements and an event process/dropout.

The level of information gained by using joint models when analysing both the longitudinal and time-to-event elements of a model, as well as details of the relationship between

them, would be far more complicated to obtain by using separate analyses.

The main focus of the thesis will be the random slope and intercept latent Gaussian variable classification of the joint model as introduced by Henderson et al, 2000 [10].

1.4.3 Model Specification

To form this joint model, the longitudinal and time-to-event elements are specified separately and linked via a latent Gaussian process. For this thesis, we specify the basic longitudinal element as

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + Z_{ij}. \tag{1.1}$$

In this model, $\mu_i(t_{ij})$ is the mean response defined as $\mu_i(t) = x_{1i}(t)'\beta_{\mathbf{1}}$, where x_{1i} represent a set of explanatory variables, $\beta_{\mathbf{1}}$ is the corresponding regression coefficients and W_{1i} is a latent Gaussian variable. The $Z_{ij} \sim N(0, \sigma_z^2)$ components are the measurement errors. Y_{ij} is the set of longitudinal measurements for patient i at their jth timepoint.

The event-time model is defined as

$$\lambda_i(t) = \lambda_0(t) \exp\{x_{2i}(t)'\beta_{\mathbf{2}} + W_{2i}(t)\} \tag{1.2}$$

where $\lambda_0(t)$ is the baseline hazard, x_{2i} is the set of explanatory variables modelled in the to time-to-event component with corresponding regression coefficients $\beta_{\mathbf{2}}$ and W_{2i} is a latent gaussian variable. In this part of the model, x_{2i} and $\beta_{\mathbf{2}}$ may have elements in common with $x_{1i}(t)$ and $\beta_{\mathbf{2}}$, but this is not a requirement. For this thesis, we generally fit the model described by equations (1.1) and (1.2) such that $\beta_{\mathbf{1}} = \{\alpha, \beta_0, \beta_1\}$, where α is a time parameter, β_0 is the intercept and β_1 is the treatment effect of the longitudinal element, and $\beta_{\mathbf{2}} = \{\beta_2\}$ is the log hazard ratio between different treatment groups.

The link between these two models is the latent zero mean Gaussian process $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$, which is present in both components, and calculated for each patient i. The nature of the model allows for a wide range of definitions for the function linking the longitudinal and time-to-event elements of the model. However in this thesis we define $W_{1i}(t) = U_{1i} + U_{2i}t$ and $W_{2i}(t) = \gamma W_{1i}(t)$ where (U_{1i}, U_{2i}) are the subject specific random effects such that $U_{1i} \sim N(0, \sigma_1^2)$,$U_{2i} \sim N(0, \sigma_2^2)$ and $Cov(U_1, U_2) = \sigma_{u_1,u_2}$. By estimating the γ parameter, the relationship between longitudinal measurements and likelihood of experiencing an event can be established. A positive gamma indicates that individuals with higher longitudinal readings were more likely to experience an event. Similarly, a larger proportion of patients dropping out with a lower longitudinal value will result in a negative γ.

Estimating the γ parameter can be useful tool for assessing the relationship between these two elements of the model. Establishing a relationship between the longitudinal outcomes and event time without using joint models can be a complicated procedure. While graphical representations of this relationship can be produced, by using joint models the statistical significance of this relationship can be established, and the magnitude of the relationship can be quantified (see Chapter 5). The γ parameter can be estimated using a mathematical method known as the EM algorithm, using a one step Newton-Raphson procedure.

1.4.4 The EM Algorithm

The EM Algorithm is an iterative procedure proposed in 1977 by Dempster, Laird, Rubin [53]. By using this algorithm, maximum-likelihood estimates can be generated for data sets which have incomplete data. It can be particularly useful in the situation where latent Gaussian variables are present in a model. An advantage of this method is its flexibility in terms of dealing with a wide range of data, and also the fact that while computation can

be extensive it is a relatively simple method to comprehend.

The EM algorithm is a two stage procedure that alternates between an Expectation (E) and Maximisation (M) step. In the E-step, an expectation of the log-likelihood function is calculated for the complete data based on the observed data, and predicted values for the missing data are generated. In the M-step, parameter estimates that maximise the function generated in the E-step are calculated. This process is then repeated using the new parameter estimates until convergence is achieved [39].

In the context of a random slope and intercept joint model as proposed by Henderson et. al., the following are details of how the EM algorithm is utilised. We refer to the event-time as T. In the E-step, initially the random effects $U = (U_1, U_2)$ are treated as missing data and expected values of $h(U)$, the functions of U_1, U_2 appearing in the log of the complete data likelihood, are calculated based upon the observed data (Y, T) and current parameter estimates [10]. Wulfsohn and Tsiatis [42] show that the conditional expectations can be written as

$$E[h(U)|Y, N] = \{\int h(U)f(T|U)f(U|Y)dU\}/f(T|Y)$$

where

$$f(N|Y) = \int f(N|U)f(U|Y)dU.$$

Where $f(U|Y)$ is the distribution of the random effects, and $f(T|U)$ is each patients contribution to the event-time component of the complete-data likelihood. The integrals above are approximated using Gauss-Hermite quadrature [10] [55]. This is due to the kernel having the same form as the Gaussian distribution.

The M-Step is then carried out. The complete data log-likelihood is maximised with each function $h(U)$ replaced by its corresponding expectation. This is estimated using

a combination of a one step approximation for the event-time parameters and a Breslow estimator for the cumulative baseline hazard, while closed form estimates are available for the other parameters [57]. Greater details of how these estimates can be achieved are found in published literature [10, 42, 56].

1.5 Discussion

RCTs are said to be the best method for successfully establishing the efficacy of clinical interventions. Years of medical research has resulted in the refinement of this method, and in modern trials significant measures are taken to ensure that potential bias is minimised. An overview of RCTs has been provided in this chapter and the guidelines for a good trial design has been provided. In longitudinal trials, where multiple measurements are taken over time, patients dropping out from a clinical trial lead to an incomplete measurement schedule. If these patients are ignored or the missing data are handled incorrectly, then this could lead to attrition bias. However, joint modelling of longitudinal and time-to-event data is a method capable of monitoring the change in longitudinal outcome over time while compensating for this dropout. In this thesis the benefits of using joint modelling of longitudinal and time-to-event data as applied to trials with dropout is determined, and the effectiveness of the method is analysed. Furthermore, methodological and design elements for the random slope and intercept model introduced in 1.4.2 are explored and developed.

In Chapter 2, greater details of the consequences of mishandling missing data are outlined and some alternative techniques for accounting for the missingness are discussed. In addition, a simulation study is used to compare the successes of the various methods used, including the joint modelling approach. In Chapter 3, a comprehensive systematic review is carried out to establish how often the various missing data methods are used in practice in longitudinal trials, and to determine the extent of the problem of missing data mishandling.

Little work has been completed in the area of power and sample size calculations in joint modelling. In Chapter 4 a sample size formula is developed for the random slope and intercept joint model, and a simulation study is used to compare the power of trials that use joint models to ones which only utilise a complete case analysis. These sample size and power calculations depend on the value of γ in a trial. Our understanding of the interpretation of the γ parameter is limited, so Chapter 5 investigates the properties of the γ parameter further using simulations. In this chapter, a visualisation of the relationship between γ and the mean change in longitudinal outcome prior to dropout is presented.

In Chapter 6, software is developed in R to carry out some diagnostic procedures for this type of modelling. These methods are then applied to the framework of a real trial dataset (MAGNETIC), which has not been previously done in statistical literature. In Chapter 7, demonstrations of how to apply joint models in RCT scenarios are given. In Chapter 8 we draw conclusions from the work by presenting the reader with further ideas for research in the field of joint modelling.

Chapter 2

Methods For Handling Missing Data In Longitudinal Studies

2.1 Introduction

In order to establish a successful method of missing data handling in a given RCT, it is essential to obtain as much information as possible about potential missing data [2]. While it is impossible to account for and be aware of all the properties of non observed data in a trial, some inferences can be made [54]. In clinical trial literature, many different methods of missing data handling methods have been proposed. While some methods are applicable to a wider range of trial designs, the most suitable method for each RCT is dependent on the unique properties of that trial [23]. A method that may be perfectly acceptable for one RCT analysis may result in misleading conclusions for another [60]. In order to establish which methods for missing data handling are appropriate in a given RCT, it is necessary to categorise the missing data [54]. In this chapter, the current methods of missing data handling are presented. However, as a prelude to this information, a description of the common missing data classification is provided.

While many articles discuss the nature of missing data handling methods in longitudinal studies [4, 24, 58, 59], no simulation study has been carried out which quantitatively compares these standard methods with a joint model analysis accounting for dropout. Theoretically the advantages and drawbacks of each missing data method can be discussed, however to gain a fuller understanding it is necessary to observe the results of each method being used in practice. To establish the impact of using different missing data handling methods in a longitudinal analysis, a series of simulation studies based on MAGNETIC data [47] are carried out in Section 2.4 to determine the extent to which different missing data methods give different results.

Furthermore, an illustration of how joint modelling can be used in an RCT for handling missing data is demonstrated through the analysis of the MAGNETIC data set. This trial had a sufficient amount of dropout (26.7%) to suggest that a complete case analysis alone

may not be an accurate reflection of the true treatment effect.

2.2 Classification of Missing Data

In general, missing data within a trial can be categorised as; Missing Completely at Random (MCAR), Missing at Random (MAR) or Missing Not At Random (MNAR). [61].

2.2.1 Missing Completely At Random

Data are Missing Completely at Random (MCAR) if the missingness indicator is unrelated to any inference that can be drawn from the dataset [59], [62]. In this case, the probability of a value being missing is unrelated to both the observed values and the values that the subject would have recorded had the data been available. An example of this is when some measurements are lost by a clinician or an equipment failure has led to a patient outcome not being recorded [63]. In both instances, the patient's status is irrelevant when considering the probability of missingness. Formulaically, let r be the missingness and y_{obs}, y_{mis} be the observed and missing values within a dataset respectively. In this thesis y_{obs} is defined as the longitudinal values for each patient pre-dropout, and y_{mis} are the potential longitudinal readings post dropout. Missing completely at random data is defined as [64]

$$Pr(r|y_{obs}, y_{mis}) = Pr(r).$$

In MCAR, the average intervention effect is the same when an analysis is performed with just the observed values as when no missing data is present. However, this may result in a significant loss of information and wider confidence intervals if the percentage of missing data is high.

In practice it is uncommon that data are MCAR [65], and there are various issues associated with determining whether data is MCAR after a patient has dropped out. By

using simple t-tests, it is possible to test for the relationship between the values of observed covariates and subject missingness. However, if no relationship is found between any covariate value and the completion of the measurement schedule by a subject, this only confirms that the data has missing properties equivalent to MCAR, which alone is not sufficient to assume MCAR data [4]. This difficulty in establishing the MCAR mechanism highlights the importance of encouraging clinicians to record the reasons behind each patient's missing data.

2.2.2 Missing At Random

In some cases when the missingness mechanism is not MCAR, data may be classified as Missing At Random (MAR). In MAR data, the missingness is dependent on observed values within the dataset [59, 61]. This can refer to baseline variables, covariate values, or previous longitudinal measurements taken in the trial. For example, in some clinical trials patients with a worse prognostic factor at baseline are less likely to complete the measurement schedule in a study for reasons unrelated to the outcome of interest. This is an example of MAR data. In terms of a mathematical formula, this is denoted [66];

$$Pr(r|y_{obs}, y_{mis}) = Pr(r|y_{obs}).$$

In the case of MAR, the missingness is not dependent on the outcome itself, however vital information is being lost by omitting the patients that had missing values [59] as the patients included in the analysis may have different prognostic outcomes to those which are omitted. This can lead to patients with a specific set of prognostic profiles being omitted from the analysis in some cases.

2.2.3 Missing Not At Random

If missing data is not classifiable as MCAR or MAR, the data is Missing Not At Random (MNAR), also known as non ignorable. In the MNAR mechanism, the missing values are not only dependent on the observed covariate values, but also of the missing non-observed values themselves [59, 61]. An example of a trial with MCAR data could be one which is designed to analyse the percentage of patients that have given up smoking. At the end of the trial, those patients who have stopped smoking are more likely to disclose the result of the trial than those who have failed to quit smoking. In a longitudinal framework, when data is MNAR there is a link between the values post-dropout and the patient dropping out. In terms of a mathematical formula, this denoted as [67]

$$Pr(r|y_{obs}, y_{mis}) = Pr(r|y_{obs}, y_{mis}).$$

For a trial with MNAR data some standard modelling methods of missing data handling may be inefficient, and more sophisticated methods of analysis need to be employed as a simple analysis may fail to compensate for important prognostic details [4].

This diversity of potential missing data should further emphasise the importance of good trial design. While in recent years many techniques of missing data handling have now been established, one of the initial aims within a trial should be to ensure that the amount of missing data is minimised [60]. Furthermore, in many cases it is difficult to classify the missingness within a trial, although this can be estimated with higher levels of clinical information [23]. To ensure that a mishandling of missing data is avoided, techniques should be discussed when the trial protocol is being constructed, as this will lead to a more transparent statistical analysis.

2.3 Methods for Missing Data Handling: Dropout in Longitudinal Studies

Longitudinal trials have an additional benefit when adjusting for missing outcome data, as measurements recorded prior to dropout for those patients can be used to enhance the accuracy of treatment effect estimates [68]. Additionally, the individual complete patient profiles for the outcome can be used as a predictive mechanism for those who did not complete the measurement schedule. While intermittent missing data in RCTs can also prove problematic, the focus of this thesis will be on missing data caused by patient dropout. In this section, an overview of the methods of missing data handling used in RCTs with longitudinal outcomes are presented, including methods of case deletion, imputation methods and modelling methods. These methods are demonstrated using data from the MAGNETIC trial.

2.3.1 Case Deletion

Case deletion is defined as the exclusion from an analysis of some individuals based on a chosen property; which may be a prognostic factor or a feature of the data for each patient. In the scenario where all patients that dropped out or had missing values are excluded, this is known as a complete case analysis [4, 78]. This is based upon the assumption that patients with missing values represent a random sample of the complete dataset (MCAR). As an example, Figure 2.1 shows a potential longitudinal dataset:

	Time (minutes)						
Patient Number	0	20	40	60	120	180	240
1	3	4	7	6			
2	8	3	1	4	6	2	3
3	1	5		6	5	6	3
4	3	4	7	6	9	8	8

Figure 2.1: Example Full dataset

A complete case analyses would result in following data being analysed,

	Time (minutes)						
Patient Number	0	20	40	60	120	180	240
2	8	3	1	4	6	2	3
4	3	4	7	6	9	8	8

Figure 2.2: Complete Cases of Figure 2.1

where patients 1 and 3 excluded. In datasets with a low percentage of patients with missing data, a complete case analysis may provide an accurate approximation to the actual treatment effect [71]. However this method is less appropriate in trials where MCAR is not a valid assumption. In particular, when higher percentages of missingness are present, this method can lead to incorrect conclusions being drawn; for example in a trial where a large number of patients drop out due to poor prognosis [73]. Also, the loss of information when using a complete case analysis will have a substantial effect on the power of the trial and confidence intervals if the percentage of missing data is too high [4, 74].

2.3.2 Imputation Methods

In order to address the problems associated with case deletion, missing values can be imputed prior to analysis. This involves substituting the missing values in a dataset for predicted values based upon the observed data [4]. Data is then analysed using standard methods. Imputation methods can be categorised as simple or multiple.

Simple Imputation Methods

In simple imputation, missing values are replaced with a single imputed value to generate a full dataset [4]. This dataset is then analysed using standard techniques. Some methods of simple imputation are outlined below.

Last Observation Carried Forward The last observation carried forward (LOCF) method takes the last observed value of a patient prior to dropout and imputes it for the rest of the values until the end of follow up [4]. While this method may be appropriate in the case of a chronic longitudinal outcome, in general LOCF fails to take into account individual patient variation post-dropout, as the imputed values are only dependent on the last available longitudinal reading. Therefore, the individual reasons for dropout are not taken into account. Figure 2.3 is an example illustrating LOCF, where values are imputed at 120-240 minutes.

	Time (minutes)						
	0	20	40	60	120	180	240
Observed	3	4	7	6			
Imputed	3	4	7	6	6	6	6

Figure 2.3: Last Observation Carried Forward

In RCT analysis, the aim is usually to find an overall distribution for the data rather than focusing on individual patient profiles, and therefore using LOCF can lead to distorted results and an underestimation of variance for an outcome when applied to an entire dataset. [4, 75, 77]. However, if the majority of patients within a trial are at a steady state prior to dropout or if a patients condition is believed to have stabilised post dropout, then LOCF may be viable. Past literature suggests that LOCF is a commonly used method despite its lack of flexibility to many different trial properties [78]. If data is MCAR, the LOCF method would be inappropriate as this would induce bias [4]. A variation of the LOCF method is the final observation carried backwards (FOCB) approach.

Best/Worst Value Imputation Unlike LOCF, some forms of imputation take individual reasons for dropout into consideration. In best value imputation, patients believed to have dropped out of the study due to good prognosis will have their best observed longitudinal reading imputed in place of the missing data. [78, 79]. The flexibility of this method allows it to be used both for patients who felt they were recovering at the time of dropout

and those that believe they have been cured completely.

Likewise if patients dropped out of the study due to poor prognosis, their previously observed worst prognostic value is imputed. In order to ensure the correct justifications for these imputation, the reasons for dropout should be verified within the study by a clinician. In some trials a combination of the best and worst case scenario is used [79]. Figure 2.4 demonstrates this imputation method:

	Time (minutes)						
	0	20	40	60	120	180	240
Observed	3	4	7	6			
Best	3	4	7	6	3	3	3
Worst	3	4	7	6	7	7	7

Figure 2.4: Best/Worst Value Imputation

The issues with using this method are similar to that of LOCF, where its appropriateness is based upon the specific prognostic factor in question. Without clinical justification, this method can also provide confusing or misleading results due to the effect on variance that excessive imputation of best and worst values can cause.

Mean Value Imputation Mean value imputation can be implemented in one of two ways; either by imputing the mean value of the patient's longitudinal outcomes for the missing values post-dropout, or by imputing the mean value of the longitudinal outcome for each timepoint [4]. This method takes into account longitudinal outcome values prior to dropout, but not the trend of each subject individually. Figure 2.5 illustrates the first procedure of mean value imputation:

	Time (minutes)						
	0	20	40	60	120	180	240
Observed	3	4	7	6			
Imputed	3	4	7	6	5	5	5

Figure 2.5: Mean Value Imputation using the mean of an individual's observed outcomes

The first type of mean imputation results in the missing outcomes having no relation to any other covariate value for each patient. This will result in a reduction in the variance estimates if the level of missing outcome data is high, or the data collected is multi-dimensional. For a longitudinal analysis with a particular interest in the change in outcome over time, this method may not be the most informative when applied to a full dataset. In the case of discrete data where the mean imputed values do not result in integer imputation, a decision must be made as to whether the values are rounded to the nearest whole number or the mean is simply imputed.

Maximum/Minimum Value Imputation Maximum and Minimum value imputation uses a similar mechanism as best and worst value imputation. After an individual has dropped out from a study, either the maximum or minimum possible value is imputed post dropout, depending on the dropout reason. For example, in the MAGNETIC trial, "9" would be imputed for patients who dropped out due to poor status and "0" for patients who dropped out due to good status, as these were the highest and lowest possible values of the asthma severity [37]. This method can be used with data that has bounded outcomes, as demonstrated in Figure 2.6. One patient in this figure dropped out due to poor status, and another due to good status with a longitudinal outcome in the range of 0 to 9.

	Time (minutes)						
	0	20	40	60	120	180	240
Observed	3	5	7	8			
Imputed by max	3	5	7	8	9	9	9

	Time (minutes)						
	0	20	40	60	120	180	240
Observed	3	2	2	1			
Imputed by min	3	2	2	1	0	0	0

Figure 2.6: Maximum/Minimum Value Imputation

As we would expect, the issues that arise when using this form of imputation are similar to those mentioned for best/worst value imputation. Variance estimates can become skewed when a large percentage of patients have values of one type imputed. Treatment effect can also be overestimated/underestimated if the number of dropouts due to positive/negative status is disproportionate between treatment groups. For example, if a large number of patients drop out for negative reasons in both treatment groups, this will result in an underestimation of treatment effect as patients these dropouts will all have the same values imputed. Likewise, the opposite is true if the reasons for dropout were different in each group, as the extreme positive and negative values will be imputed.

Multiple Imputation

Multiple imputation is a missing data handling technique conceived by Rubin in 1987 [80]. It involves replacing missing values with at least two potentially feasible sets of values. This offers a more flexible method of missing data analysis than simple imputation methods and also manages to address the issue of variability which is not covered by simple methods [80]. Simple imputation methods generally overestimate the level of precision as there is an omission of the between imputation variability component [81]. In order to compensate for the uncertainty of missing data, multiple datasets can be generated and the results pooled to give more accurate estimates [7]. As with simple imputation methods, multiple imputation methods when poorly planned are likely to induce misleading results.

While the concept of multiple imputation may not seem too appealing to the clinician at times, as some may feel it is merely creating data, the subject is mature and rigorous in it's development. With the technological advances leading to a greater power of computational software in the past few years, the generation of multiple imputed datasets has become a more manageable procedure, with software such as R [46], SAS [76] and WINMICE all having multiple imputation capabilities.

MICE is a commonly used variation of multiple imputation, as developed originally by Van Buuren, Oudshoorn in 1999 [82]. It utilises an imputation algorithm which stochastically draws from a distribution generated from the observed values of each variable in order to provide multiple approximations for the missing data.

The MICE algorithm works as follows, [82], let $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$ be the set of random variables in a study, which can include outcome variables. All these variables can have missing data. These are made up of the observed and missing components, $X_j = (X_j^{obs}, X_j^{mis})$.

To impute, draw from the unconditional multivariate distribution of X, by assuming the data is missing at random. The algorithm generates the estimates as follows:

- Set the counter $a = 1$.
- Draw values of $X_1^{mis,2}$ from $P(X_1|X_2^1, X_3^1, \ldots, X_k^1)$.
- Draw values of $X_2^{mis,2}$ from $P(X_2|X_1^2, X_3^1, \ldots, X_k^1)$.

$\vdots$

- Draw values of $X_k^{mis,2}$ from $P(X_k|X_1^2, X_2^2, \ldots, X_{k-1}^2)$.

$\vdots$

- Draw values of $X_k^{mis,a+1}$ from $P(X_k|X_1^{a+1}, X_2^{a+1}, \ldots, X_{k-1}^{a+1})$.

$\vdots$

This procedure continues until the values converge. A common occurrence is that this algorithm is repeated 5 times to create separate imputed datasets. Applying the MICE algorithm in practice requires the conditional distributions, from which the missing data can be drawn, for each variable to be specified. These distributions are easily specified when using linear modelling techniques. For the purpose of using MICE in a longitudinal framework, the outcome measurements taken at each time point, $Y_{i1}, Y_{i2}, \ldots$, can be treated as separate variables. Each completed dataset is then analysed, with the estimates pooled using standard methods. As the number of imputations in MICE tends to infinity, the imputed complete model corresponds to a mixed model for linear regression based imputation [85].

The flexible nature of this algorithm means that it is not limited to imputing only continuous variables. For example, if a binary variable is being imputed then these values can be drawn from a logistic regression model. Similarly there are many methods of drawing from distributions for ordered categorical variables and discrete variables [86].

2.3.3 Modelling Methods

Linear Mixed Models

An alternative to case omission or imputation methods is simply by fitting linear mixed models to the observed values in a dataset [87]. These models are a sufficient method of analysis in the case where the missingness mechanism is MAR, or missing data and dropout percentages are low [87].

In this thesis, the focus will be on linear mixed effects models with a random slope and

intercept, which has the following form:

$$Y_{ij} = x_{1i}(t)'\beta_{\mathbf{1}} + b_{1i} + b_{2i}t + \epsilon_{ij} \tag{2.1}$$

$$b_{i1} \sim N(0, \sigma^2_{z1}), b_{i2} \sim N(0, \sigma^2_{z2}), Cov(b_{i1}, b_{i2}) = \sigma_{z12}$$

$$\epsilon_{ij} \sim N(0, \sigma^2\omega_{ij}).$$

In formula 2.1, the fixed effects are denoted by $\beta_{\mathbf{1}}$, and the random effects by b_{1i} and b_{2i}. In general $\beta_{\mathbf{1}}$ is analysis specific, however for this thesis, as in Section 1.4.3 we define $\beta_{\mathbf{1}} = \{\alpha, \beta_0, \beta_1\}$, where α is a time parameter, β_0 is the intercept and β_1 is the treatment effect of the longitudinal element.

Joint Modelling For Time-to-dropout

In Chapter 1 the concept of joint modelling of longitudinal and time to event data was introduced in a general context. One advantage of this flexible method is that a longitudinal outcome can also be modelled while accounting for dropout. In this case event time is replaced by dropout time and statistical inferences can be made such that;

- Inferences about the longitudinal measurements being adjusted to compensate for potential informative dropout based on treatment group.
- The distribution of dropout while compensating for longitudinal measurements.
- Analysis of how the model evolves with respect to the longitudinal measurements and dropout.

The formulation specified by Henderson 2000 [10] can be used to jointly model a longitudinal outcome and dropout as described in Chapter 1, Section 4.3.

2.3.4 Other Methods

As well as the imputation and modelling methods described in this chapter, there are alternative ways of taking into account dropout that can be combined with any of the above. Two of the most common are the Weighted Least Squares Method (WLSM) and Sensitivity Analyses.

In the WLSM, weights are assigned to the complete cases with the purpose of reducing bias. All patients are allocated into subgroups based on their baseline readings and response profiles so that each group contains patients with similar prognostic profiles regardless of whether a patient completed the study. Prior to analysis, the individuals who complete the study despite having a high probability of dropout can have a large influence on the results using a weighting system. Greater details of how WLSM is utilised, as well as the various proposed methods of weighting can be found in published literature. [83, 84]

Alternatively, a sensitivity analysis assumes an uncertainty in the trajectory of patients post-dropout. In a sensitivity analysis, a number of potential different scenarios are considered post dropout, and imputations are made based on these scenarios. The multiple generated datasets are then analysed separately to determine whether the same conclusions are reached in the different analyses, and the plausibility of each imputation is discussed. An example of this in an RCT setting would to by generate the following 3 separate data sets; 1) Patients that dropped out have the worst case value imputed post-dropout, 2) Patients that dropped out have the best case value imputed post-dropout, 3) Missing values are imputed using LOCF. By observing the results from all 3 separately an uncertainty analysis can be utilised to provoke a discussion of the overall efficacy of a treatment.

2.4 Simulation Study

Throughout the chapter, emphasis has been made on the importance of choosing the correct method for missing data handling. Each choice of missing data handling method should be discussed and justified extensively prior to analysis as an incorrect choice of imputation method can lead to incorrect assessments of treatment effect [69]. While the optimal missing data handling method in an RCT is trial specific, some methods may be more appropriate than others across a wider range of trial properties.

In the past, many reviews of missing data techniques have been conducted, although very few of these discuss techniques in longitudinal trials exclusively. Of those that do, no studies compare different imputation methods alongside joint modelling methods by using real data to generate simulated datasets in a longitudinal environment. By applying imputation methods to simulated datasets and comparing the results, it is possible to establish the advantages and drawbacks of using each imputation method and look at how successful each method is in estimating treatment effect. In this section, a simulation study based on the parameters of the MAGNETIC dataset is presented. This is something which has not previously been published in statistical literature.

2.4.1 Simulating Data

Mimicking the MAGNETIC trial scenario, joint longitudinal and time-to-dropout data are generated. The aim was to generate data for patients randomised to one of two treatments, with four different percentages of dropout; 20%, 30%, 40% and 50%. By simulating data using joint modelling methods, a higher probability of dropout can be given to patients with a higher longitudinal profile which ensures that the data is not MCAR. 1000 datasets were simulated for each percentage of dropout. Various imputation methods were then employed to each data set, and separate analyses were carried out using linear mixed mod-

els. The results were collected to establish how accurate each method was for estimating the true treatment effect β_1.

Outcome data was simulated from a longitudinal mixed model with variables of time (α), intercept (β_0) and treatment (β_1) (see Section 1.4.3). Longitudinal timepoints were set at 0, 60, 120, 180 and 240 hours in order to mirror the MAGNETIC follow-up time. Event-time was simulated from the Cox proportional hazards model with a flat baseline hazard, also based on MAGNETIC parameters [112]. This was then transformed into dropout time, defined as the last time point prior to the event-time generated for each subject. To simulate data that mirrored the MAGNETIC trial (which will be demonstrated in Section 2.5), we fixed $\alpha = -0.0076$, $\beta_0 = 5.617$, $\beta_1 = 0.2$ and $\gamma = 0.18$. $\beta_2 = -0.5$ and the variances of (U_1, U_2) were set to 1.19 and 0.00003 respectively. The covariance between the random effects was set to 0 and the error variance was set at 0.5.

For these simulations, continuous outcome data was generated, as opposed to discrete as in Magnetic, as we were interested in establishing the success of each method when modelling a continuous longitudinal outcome. However, as one of the missing data handling methods to be utilised was highest/lowest value imputation as employed in the MAGNETIC dataset, it was desirable for the longitudinal outcome in the simulations to be within the range of 0 to 9. To achieve this range, values < 0 or > 9 were rounded to 0 or 9. To ensure that this modification had a negligible effect on the true treatment effect in the analysis, 10000 preliminary datasets were generated and the percentage of outcomes not in the desirable range was calculated. The average percentage of longitudinal values not in the range of 0 to 9 was 0.072%. This was considered an acceptably low value, and would have a negligible effect on the simulated results.

2.4.2 Methods

For each method of analysis and missing data handling technique, one of two models were fit to the data. For complete cases and all imputation based methods, a standard longitudinal linear mixed model was fitted with random slope and random intercept. The form of this model is shown in Equation (2.1), with $\beta_{\mathbf{1}} = \{\alpha, \beta_0, \beta_1\}$ included as fixed effects.

For the joint modelling analysis, a random slope and intercept joint model was fit to the data as described in Equations (1.1) and (1.2) in Chapter 1, Section 4.3. Similarly, we defined $\beta_{\mathbf{1}} = \{\alpha, \beta_0, \beta_1\}$ and $\beta_{\mathbf{2}} = \beta_2$ to model the hazard ratio between treatment groups. The following descriptions detail how the missing data methods were applied:

- **Complete Cases** - Only patients with no missing data are included and a linear mixed model was fit to estimate treatment effect. .
- **LOCF** - Patients that dropped out had their last longitudinal value repeated until the end of the study. The dataset was then modelled using the same linear mixed model.
- **High/Low Imputation** - Patients with a longitudinal outcome of 4.5 or higher at dropout had the value of 9 imputed for the remainder of the study, and likewise those lower than 4.5 had 0 imputed. The data was analysed using linear mixed models.
- **Mean Imputation** - The mean value of the patients observed values were imputed for patients that dropped out, and linear mixed models were fit.
- **MICE** - The missing values were multiply imputed 5 times via mixed modelling techniques. Analysis was carried out using linear mixed models.
- **Mixed Models** - Only the observed values were fit using the linear mixed model.
- **Joint Modelling** - The parameters were estimated using the E.M. algorithm. Details of how the joint model are fit have been provided previously in section 2.4.2.

For the majority of the methods above, ready made code was available in R to carry out the imputations. However, this was not the case for high/low imputation, so a novel function was generated in R to employ these methods (see Appendix). To establish the success of each method the mean estimate of β_1, the mean square error (MSE), the mean standard error and relative bias are calculated for β_1. To calculate the standard error of the parameter β_1, standard methods are used for the linear models, and bootstrapping methods are applied in the case of joint modelling. MSE is defined as

$$MSE = \frac{\sum_{i=1}^{1000} (\hat{\beta}_{i1} - \beta_1)^2}{1000}$$

as data was simulated 1000 times in each setting, and i is the simulation number. Relative bias is defined as the mean percentage of error in estimation of β_1, with the mathematical formula:

$$\left| \frac{\beta_1 - \bar{\beta}_{i1}}{\beta_1} \right| \times 100.$$

where $\bar{\beta}_{i1}$ is the mean estimate of β_1 across the simulations. The most successful methods of estimation were those with low mean square errors and mean estimated treatment effects close to the true value.

2.4.3 Results and Discussion

Table 2.1 displays the results for the simulations with different percentages of dropout. The true value of β_1 was 0.2. Figure 2.7 graphically shows the mean estimates of β_1 and the MSE of the treatment effects for each missing data method.

Dropout	Method	Mean Estimate $\hat{\beta}_1$	MSE	Standard Err.	Relative Bias(%)
20%	Complete Cases	0.2268	0.01147	0.1066	13.4
	LOCF	0.2108	0.00967	0.0909	5.4
	High/Low	0.2064	0.00931	0.0885	3.2
	Mean Imp	0.2012	0.00944	0.0939	0.6
	MICE	0.2028	0.00942	0.0927	1.4
	Mixed Models	0.2030	0.00942	0.0921	1.5
	Joint Modelling	0.2030	0.00941	0.0918	1.5
30%	Complete Cases	0.2620	0.01852	0.1218	31.0
	LOCF	0.2176	0.00983	0.0924	8.8
	High/Low	0.2059	0.00911	0.0889	3.0
	Mean Imp	0.2026	0.00994	0.0941	1.3
	MICE	0.2040	0.00969	0.1054	2.0
	Mixed Models	0.2048	0.00968	0.0932	2.4
	Joint Modelling	0.2041	0.00966	0.0923	2.1
40%	Complete Cases	0.2679	0.02105	0.1410	34.0
	LOCF	0.2148	0.00992	0.0932	7.4
	High/Low	0.1886	0.00860	0.0932	5.7
	Mean Imp	0.1977	0.01016	0.0949	1.2
	MICE	0.1970	0.00998	0.0935	1.5
	Mixed Models	0.2031	0.01004	0.0951	1.6
	Joint Modelling	0.2009	0.00992	0.0933	0.5
50%	Complete Cases	0.2701	0.02326	0.1685	35.1
	LOCF	0.2242	0.01031	0.0979	12.1
	High/Low	0.1792	0.00851	0.0939	10.9
	Mean Imp	0.2040	0.01035	0.1009	2.0
	MICE	0.2016	0.01009	0.1002	0.8
	Mixed Models	0.2046	0.01030	0.0974	2.3
	Joint Modelling	0.1967	0.00995	0.0965	1.7

Table 2.1: Simulation Study Results

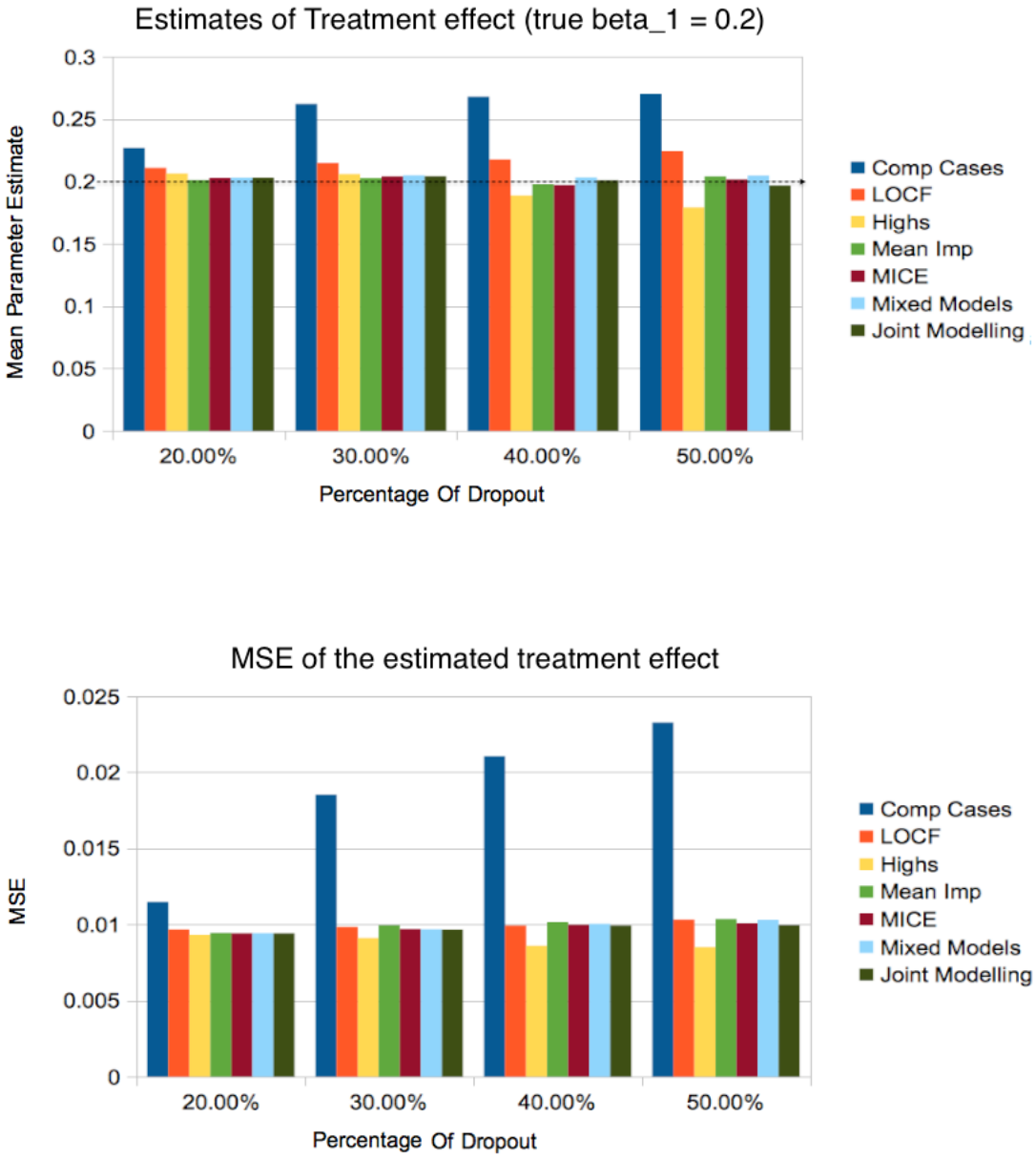


Figure 2.7: Mean treatment estimate and MSE for each missing data method

When dropout was set to 20%, complete case analysis and LOCF methods yielded the greatest overestimation of true treatment effect. Similarly, the method of complete case analysis has the highest MSE, indicating that this method performed the worst. The lowest MSE resulted from a combination of highest/lowest value imputation, while mean imputation methods displayed a mean treatment effect estimate closest to the actual value. Joint modelling methods performed well, with the second lowest MSE and a relative bias of 1.5%. There did not appear to be a large variation in estimate of $\hat{\beta}_1$ for each method.

In the simulations with 30% dropout, the magnitude of overestimation of treatment effect was far greater in the complete cases than the previous simulation. The MSE was also higher for the complete case analysis than the alternative missing data handling methods. In general the MSE's were higher than for the previous simulation, with the exception of highest/lowest imputation, which had the lowest MSE. A mean imputation analysis yielded the β_1 estimate closest to the true value, while MICE, mixed models and joint modelling methods also gave low relative biases.

In general, the high/low imputation method underestimated treatment effect more than the other methods in the simulation with 40% dropout, although this MSE was again the lowest for these imputations. Joint modelling methods performed well, yielding the lowest relative bias and a low MSE. β_1 in a complete case analysis was on average overestimated by 34.0%. LOCF methods resulted in higher relative biases than the other imputation methods.

With a 50% dropout a greater range of results would be expected for the different methods of missing data handling. However, in these simulations, MICE provided treatment effect estimates closest to the true value. Mixed models, joint modelling and mean imputation methods also still performed well despite the high dropout percentage. When observing the relative bias; a complete case analysis, LOCF and a combination of high and

low imputation had considerably higher values than the other methods which may shows that they performed poorly with such a high percentage of dropout.

When comparing the MSE and mean parameter estimate to the actual treatment effect (0.2), it can be observed that in general the high/low imputation methods, as used in MAGNETIC, appeared to perform successfully with the lowest MSE across the 4 simulation studies. However, in general the high/low imputation method underestimated the treatment effect for higher percentages of missing data which led to higher values of relative bias. The slight overestimation of treatment effect in general may be caused by different percentages of patients dropping out in the treatment groups, and the fact that patients with higher profiles were more likely to dropout. Despite the common use of the technique in many trials, complete cases as expected performed the worst, with LOCF also performing comparatively poorly. With 30% dropout and higher, the mean estimation of treatment effect for a complete case analysis was a more than 30% overestimation of the actual value.

Across all simulations joint models, MICE, mixed modelling methods and mean imputation methods were the most consistent at estimating β_1.

2.5 Application to MAGNETIC Data

As previously highlighted, the motivating dataset for this thesis came from the MAGNETIC trial which was introduced in detail in Chapter 1. Overall 26.7% of patients dropped out from this trial, so a complete case analysis alone may not provide accurate results of treatment effect. In the primary analysis for MAGNETIC, ASS was found to be significantly lower for patients randomised to magnesium at $t = 60$. In this section, a joint modelling analysis is applied to MAGNETIC, which allows inferences to be made about ASS, dropout and the relationship between these components.

2.5.1 Exploratory Analysis

Figure 2.8 shows the mean longitudinal profiles of the ASS for the two treatment groups.

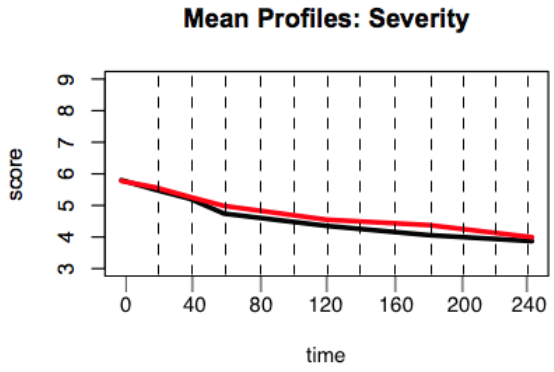


Figure 2.8: Mean Profiles of Asthma Severity Score for Magnesium (Black line), Placebo (Red line)

The mean ASS was approximately the same between the treatment groups at baseline (5.72 in the magnesium group, 5.75 in the placebo group), however the mean ASS became marginally lower for patients in the magnesium group as time progressed. At the original primary outcome of interest, $t = 60$, the severity score was 4.72 in the magnesium group and 4.95 for patients randomised to placebo.

Figure 2.9 shows plots of the mean longitudinal profiles for patients that dropped out at each time point, by treatment group.

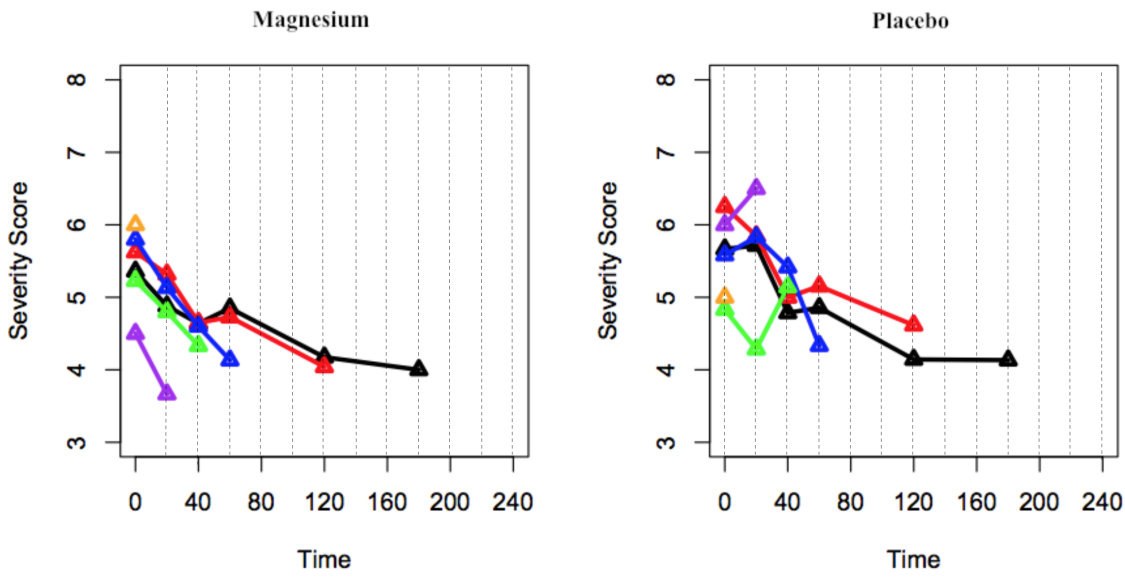


Figure 2.9: Mean Severity Score for Dropouts at each time point

Overall, the mean profiles showed that most patients had a decreasing ASS pre-dropout, which indicates that the data is not MCAR. The mean ASS of patients randomised to magnesium decreased prior to dropout for all time points, and magnesium randomised patients that dropped out earlier in the study had a greater mean decrease in ASS. However, in the placebo group, patients that dropped out at $t = 20$ and $t = 40$ had an increasing mean ASS pre-dropout. This may be informative as treatment stopped being administered in this study at $t = 60$. These results indicate that patients with a positive prognosis were more likely to drop out.

2.5.2 Estimating the treatment effect

To investigate the treatment effect in the longitudinal context, β_1, statistical models were fit to the data and four separate methods of analysis were conducted; complete case, maximum/minimum imputation, MICE and joint modelling. For imputation methods and

complete case analysis, the random slope and intercept linear mixed model from Equation (2.1) is fit, with the same fixed and random effects parameters as described in Section 2.4.2. For maximum/minimum imputation, information was available about the reasons for dropout from the trialists involved in this study. "9" was imputed for patients post-dropout that withdrew due to poor prognosis, and likewise "0" was imputed for withdrawals due to good prognosis, while patients where no details of reasons for dropout were given are omitted from the maximum/minimum analysis

For the joint modelling analysis, the random slope and intercept model described was fit as described in Section 2.4.2. In the dropout element, treatment effect β_2 was modelled using the Cox-Proportional Hazards model. The results are presented in Table 2.2.

Model	Parameter Estimate	Std.Error	p-value
Complete Case - β_1	-0.1293	0.1112	0.2452
Maximum/minimum Imputation - β_1	-0.1601	0.1107	0.1488
MICE - β_1	-0.2132	0.1049	0.0332
Joint Modelling - β_1	-0.2048	0.1041	0.0408
Joint Modelling - β_2	0.5533	0.1789	0.0103
Joint Modelling - γ	-0.1823	0.0990	0.0318

Table 2.2: Analysis of MAGNETIC ASS

In the analysis, a complete case analysis and maximum/minimum imputation analysis did not detect a significant difference between the two treatments, which is not the case when MICE and joint models are used. In both analyses with case deletion, the differences between the treatments was underestimated when compared to MICE and joint modelling. Using joint modelling methods, it was determined that patients randomised to magnesium had an ASS that was 0.20 points lower than in the placebo group, which was a statistically significant outcome. The log hazard ratio was estimated at 0.55, and therefore patients randomised to magnesium were more likely to leave the study. This appears to concur with Table 1.1 in Chapter 1. γ was also found to be significant and negative, which indicates that

patients with a lower severity score were more likely to drop out, which is demonstrated in Figure 2.9.

2.6 Discussion

In the simulation study, the joint model performed consistently well for all percentages of dropout, with the MSE's being low and the mean treatment effect being approximately the same across the simulations when compared with the other missing data methods. This demonstrated that joint modelling can prove to be a useful tool when analysing longitudinal data while compensating for dropout. Some missing data methods resulted in inaccurate treatment effect estimates for larger percentages of missingness. The study indicated that a complete case analysis will often yield skewed or inaccurate results when non MCAR data is present, particularly for higher percentages of missingness.

The diversity within the results of simulation study reinforces the importance of carefully selecting an appropriate missing data method in a trial. One advantage of conducting a sensitivity analysis, in which multiple imputation mechanisms are applied to the post-dropout data, is that the merits of each imputation method can be discussed and conclusions drawn from the multiple analyses. By presenting the information from a sensitivity analysis in an understandable way, the clinician is able to utilise the information to provide more in depth results.

The MAGNETIC trial provided an opportunity to illustrate the joint modelling framework in a trial with longitudinal data and a high percentage of dropout. The joint model fit to the data estimated that patients randomised to magnesium had an ASS of 0.2048 points lower than in placebo, which was a statistically significant difference. However, due to the small magnitude, the clinical significance of this result is something to be discussed by trialists and physicians. The estimate of treatment effect was found to be underestimated

by using a complete case analysis, and a statistically significant difference was not detected.

Children randomised to magnesium were also found to be more likely to dropout from the study. By modelling γ, the negative estimate shows that patients were more likely to leave the study if they had a lower severity score. This seems to indicate that children may have been withdrawing from the study more frequently due to good status/the feeling that their asthma symptoms had improved. A more detailed discussion of γ interpretation is presented in Chapter 5.

The simulation study results provide further evidence of the benefits of using joint modelling analyses in clinical trials, as this method performed as well as imputation analyses and mixed models. MAGNETIC was a perfect illustration of a trial which benefited from an additional joint modelling analysis. By fitting this type of model, information was gained which would have been more difficult to obtain by carrying out separate analyses.

In terms of the statistical literature, joint modelling is still a relatively new methodology. Despite the advantages of using joint modelling some clinicians or statisticians may be unaware of how to apply the model, and therefore it may be under-utilised in trials which model a longitudinal outcome while some patients drop out. In the next chapter a systematic review is carried out to establish just how often joint modelling and the other missing data handling methods are used in trials with longitudinal measurements.

Chapter 3

A Review Of The Handling Of Missing Longitudinal Outcome Data in Clinical Trials

3.1 Introduction

In the previous chapter, methods were presented that aim to counteract the problem of missing data in RCTs with longitudinal measurements. In this chapter an investigation is carried out to establish how often these methods are used in practice. The problem of handling missing data is frequently highlighted in RCT literature [92]. In well designed trials, measures are taken to ensure that the level of missingness is as low as possible, however patient dropout is inevitable in most RCTs and high percentages of missing data can prove problematic in establishing the true clinical effectiveness of a treatment [71]. Since the last review of missing data handling techniques published by Wood et al. in 2005 [89], there has been significant development and progress in the area of missing data handling. In Chapter 3, a systematic review is conducted to observe which missing data handling methods are used in trials with longitudinal outcome data, published between 2005 and 2012. The work in this chapter has been published in the Journal of Trials [60], and this article has been used to influence the missing data handling methods used in the PEPTIDE randomised control trial [72].

By reviewing all trials published between July 2001 and September 2001 in four leading journals, White, Wood and Thompson came to the conclusion that missing outcome data was an issue that has failed to be addressed in an appropriate manner in many RCTs, and that missing data was "often inadequately handled". Of the papers in the study with longitudinal outcomes, 17 out of the 37 (46%) used a complete case analysis to deal with the missing data, which can yield misleading conclusions as demonstrated in Chapter 2. Of the RCTs with missing longitudinal outcome data included in the study, 7 out of 37 (19%) papers used imputation methods. [89].

However, the majority of papers in the aforementioned review did not contain longitudinal outcome data, and little research has been done into how missing data is handled

specifically in RCTs with longitudinal outcomes. As establishing the reasons for missingness within a trial and using appropriate methods based on this information is of utmost importance, one of the primary focuses of this chapter is to assess the level of information detailed about missing data within recently published RCT articles. Unlike previously published systematic reviews of missing data handling methods, this chapter exclusively focuses on trials with longitudinal outcome data. The results in this chapter motivate the proposal of a four point plan in Section 3.4.2, which aims to provide a rigid guideline for missing data handling in future RCTs.

3.2 Methods

This systematic review aims to establish how often the trialists reported the reasons for missing data, and how the missing data was dealt with in the analysis. In particular, details are extracted about the use of imputation methods in these trials; whether imputation was used, and if an explanation for doing so was provided in the text.

Inclusion/Exclusion Criteria

Only papers that contained balanced repeated measures outcomes, and are described as a "randomised control trial" in the abstract were included in this study. Papers that were published from the years 2005 to 2012 were included and no restrictions were put on journal. All papers that were not written in English, or had non human participants as the subjects were excluded from the study, as well as any papers with only binary outcomes being recorded longitudinally, as these require different techniques of missing data handling.

Definitions and Study Selection

For this study, dropout was defined as a patient having withdrawn from the trial, with no more longitudinal readings until the end of the follow up period. While intermittent missing data is also an issue, the main focus of our extracted data will be on those individuals who withdrew from each trial.

To identify potential papers for inclusion, MEDLINE (OVID interface) was searched using the following terms; longitudinal randomi$ed controlled trial$ or repeated measure$ randomi$ed controlled trial$ or longitudinal RCT$ or the same searches with "controlled" replaced by "control". The papers identified also had to fall within the constraints of the pre-specified eligibility criteria.

Of all the papers identified as eligible for this study, it was decided at the outset that 100 would be selected at random for inclusion due to time constraints. This was done in order to ensure that there was a realistic number of papers for the time available allocated within this thesis. This randomisation was done by ordering the papers alphabetically by first authors surname, giving each of these papers a number, and then randomly generating a sequence of the integers using the "random" function in the R statistical software. The data was then extracted from each full paper in the order generated. If a paper was found to be ineligible on closer inspection then the 101st paper in the sequence was added to the study, and then the 102nd and so on. If less than 100 eligible papers were identified then all papers were included.

When the extraction was carried out, 10 of the first 100 papers read were found to be ineligible. Therefore a further 10 papers selected at random were read in full, and all these were found to be eligible. Further details of this can be found in Section 3.3 and the CONSORT diagram in Figure 3.1.

Data Extraction

For each RCT, data was extracted relating to the general characteristics of the trial, as well as more specifically the details relating to missing data handling. For the missing data handling methods, the main focus was those used for patients that dropped out. The nature of the longitudinal data; the number of longitudinal time points and whether a primary longitudinal outcome was recorded were extracted. As the need for imputation based methods is greater for larger amounts of missing data, details of the percentage of completing patients were extracted from each trial. The imputation methods used in each study were recorded, as well as the level of explanation for using the chosen method for missing data handling, and whether each trial recorded reasons for dropout within a study.

As an extra analysis, an assessment was made as to whether the methods for missing data handling used were appropriate in each paper based upon the description of the observed and missing data, the justifications provided by the authors and the percentage of missingness. In the cases where this was unclear, my first supervisor was consulted for an additional opinion.

Additionally, the presence of a statistician as one of the authors of the study and the software used for the analysis was also extracted. Finally, for the papers that used imputation, information was collected about whether a comparison was made between the complete case and imputed datasets, as well as whether these analyses yielded different statistical conclusions.

3.3 Results

Our search identified 882 articles, and abstracts were screened for potentially eligible papers. A total of 381 papers appeared to be eligible for inclusion from the abstract. After a

randomisation of the order of these papers, a further 10 of the first 100 papers were found to be ineligible due to the unbalanced nature of their longitudinal readings, after reading the full articles. A CONSORT diagram of the progress is shown in Figure 3.1.

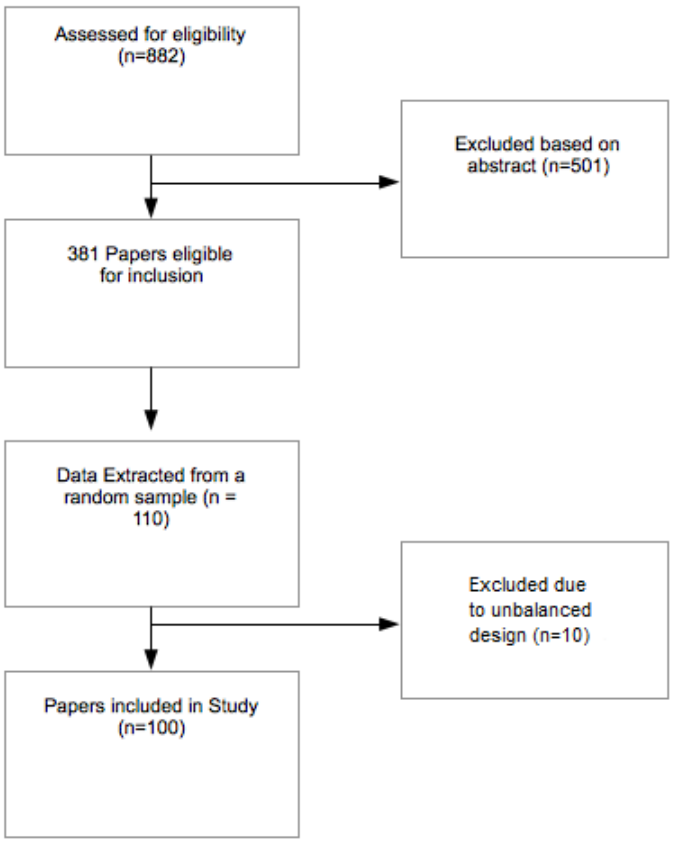


Figure 3.1: CONSORT diagram showing inclusion of papers in the study

Methods of Imputation

Data was collected from papers from a wide range of different medical areas in order to investigate how missing data was handled in practice for randomised control trials with longitudinal measurements. The most popular medical areas were mental health (13%), cancer (11%) and rheumatology (10%). Greater detail of the properties of these trials is

provided in Table 3.2, and Table 3.1 lists the primary method of imputation for missing data handling within each trial.

Primary Approach to Analysis	**Papers**
Complete Case Analysis	32 [1]
Mixed Models	18
Simple Imputation	
LOCF/FOCB/Baseline Carried Forwards	*9*
Average Value either side Imputed	*1*
Simple Algorithmic Based Imputation	*1*
Mean of other patients values imputed	*1*
Median values imputed	*1*
Multiple Imputation Methods	4
Other non-imputation based methods [2]	14
Exclusion based on amounts of missingness	6
Exclusion based on reasons for missingness	1
No missing data	9
Unclear	3

Table 3.1: Method of Missing Data Handling

Of the studies reviewed, 9 trials had no missing data (9%), and 3 papers in the "Unclear" category made no reference to missing data in their articles (3%) and therefore the level of missingness was unclear. Out of the 100 papers, 18 had used imputation (18%) and only 4 had used multiple imputation (4%). One paper carried out a complete case analysis as the primary method of missing data handling, but included an analysis based on last observation carried forward (LOCF) as a secondary method. For the papers that only included the complete cases, the most common methods of analysis were variations of ANOVA or ANCOVA in 13 trials (13%), mixed modelling in 6 trials (6%) , t-tests for mean comparison in 5 trials (5%) and linear regression modelling in 4 trials (4%).

The most common method of simple imputation used was LOCF, or a variation of

[1] One paper which used a complete case analysis, also used simple imputation as a secondary analysis. In Table 3.2, this paper is included in the simple imputation section.

[2] e.g Comparison of means, t-test, RMANOVA

this method. LOCF was used in 8 papers (8%). The clinical topics in these papers were QoL based on moral support in patients with depression, chronic muscle based neck pain, chronic arm pain due to repetitive use, shoulder pain in stroke patients, outcomes in COPD (Chronic Obstructive Pulmonary Disease), stress levels in arthritic patients, amount of sleep in patients with chronic insomnia and number of behavioural disturbances in patients with dementia. In one trial, the baseline value was carried forward to impute outcome data at 2 and 6 months for patients with missing values. LOCF may be a reasonable method if patients are in a steady state closer to the time of dropout. Of the papers that used simple imputation as a primary method of analysis, 5 trials used t-tests for comparison of means (36%), 5 used linear mixed models (36%), 3 used a variation of ANOVA (21%) and 1 used chi-squared test (7%). Of the 4 trials that used multiple imputation methods, 2 used mixed modelling, 1 used linear regression modelling and 1 used t-tests as the primary method of analysis.

			Method of Handling Missing Data							Acceptable Method? [5]		
		No. of Papers	Complete Cases	Simple [1]	Multiple	Mixed Models	No Missing Data	Unclear	Other [2]	Yes	No	Unclear
Number of Patients	1-100	48	17	8	0	4	8	3	8	18	9	13
	101-200	22	8	2	2	5	1	0	4	9	11	1
	201-300	12	4	2	0	4	0	0	2	8	4	0
	301-400	7	1	1	1	1	0	0	3	3	3	1
	400+	11	1	1	1	4	0	0	4	6	3	2
Country of Publication	USA	53	20	7	2	10	4	0	10	20	18	11
	UK	40	8	6	2	7	3	3	11	20	12	5
	Denmark	3	2	0	0	0	1	0	0	2	0	0
	Netherlands	3	1	1	0	1	0	0	0	2	0	1
	Japan	1	0	0	0	0	1	0	0	0	0	0
Number of Timepoints	3	38	13	4	3	10	1	1	6	14	16	7
	4	28	9	4	1	4	2	2	6	14	4	8
	5	16	4	3	0	3	2	0	4	10	3	1
	6	6	2	0	0	0	2	0	2	1	3	0
	7	4	1	0	0	0	2	0	1	1	1	0
	8+	8	2	3	0	1	0	0	2	4	3	1
Year	2005-06	19	9	2	0	1	2	1	4	7	4	6
	2007-08	27	8	3	0	5	4	0	7	11	10	2
	2009-10	25	7	3	1	4	3	1	6	9	8	5
	2011-12	29	7	6	3	8	0	1	4	17	8	4
Clinical Area	Mental Health	13	2	4	2	1	1	1	2	7	2	3
	Cancer	11	4	1	1	4	0	1	0	5	4	2
	Rheumatology	10	4	1	0	2	1	0	2	3	5	1
	Infectious Diseases	8	4	1	0	1	1	0	1	4	2	1
	Heart and Circulation	7	1	2	0	1	0	0	3	3	2	2
	Dentistry/Oral Health	6	3	1	0	0	2	0	0	3	1	0
	Neurology	6	2	0	0	2	0	0	2	2	2	2
	Anaesthesia and Pain	6	3	0	0	1	0	0	2	2	2	2
	Other[3]	33	8	4	1	6	4	1	9	15	10	4
Dropout Reasons Recorded?	Yes	35(39.8%) [4]	15	7	0	7	NA	NA	6	20	11	4
	Partial Information	25(28.4%)	7	5	1	7	NA	NA	5	15	9	1
	No	28(31.8%)	10	2	3	4	NA	NA	9	9	10	9

Table 3.2: Trial Characteristics

[1] One paper which used a complete case analysis, also used simple imputation as a secondary analysis. In Table 2, this paper is included in the simple imputation section.

[2] Comparison of means e.g t-test, RMANOVA

[3] A full list of medical areas is included in the appendix

[4] Disregarding the 9 papers without missing data and 3 papers where the missing data handling method was unclear

[5] Based on the 91 papers which had missing data.

Explanation of the reasons for using the statistical methods for handling missing data

It was found that 37 (42.0%) papers with missing values made comments on why they had used their particular missing data handling method of choice. The level of explanation ranged from one line statements about the efficiency of the chosen method, to multiple page descriptions of different missing data methods and the merits of each.

Of the 88 papers in which missing data was present, 51 (58.0%) provided no explanation of the reasons for the missingness methods used. All 4 papers that employed multiple imputation methods provided explanation of the reasons for their use. Out of the 14 papers that used simple imputation methods, 7 (50%) gave explanation of the reasons for their choice of imputation. 26 out of 70 (37.1%) papers without imputation explained the reasons for the missing data handling method chosen. For the papers with no missing data, 1 of the 9 (11.1%) discussed potential missing data methods with a view to suggesting how the missing data would be analysed if some had been present.

Was a statistician involved in the analysis?

Out of the 88 papers with missing data, 30 (34.1%) had a statistician cited as one of the co-authors of the study. Results indicated that there was little difference in the levels of explanation of the missing data methods used when a statistician was co-authoring a paper. 13 out of the 37 (35.1%) papers which justified their missing data methods had a statistician present, compared to 17 out of 51 (33.3%) papers which failed to explain the reasons for their chosen method.

Number of Publications by Year

It was assessed whether a greater number of papers had used imputation in recent times. Figure 3.2 shows the number of trials using imputation methods for each year.

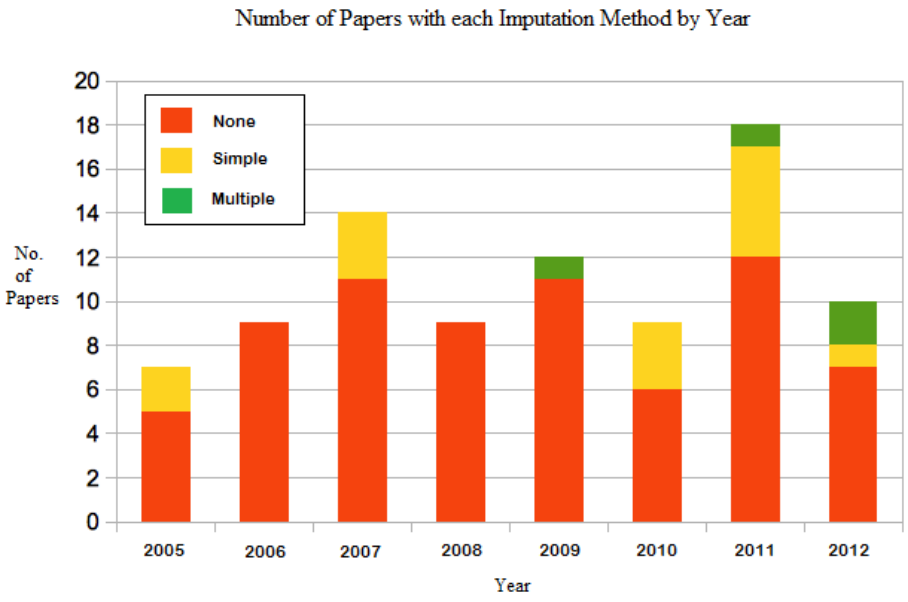


Figure 3.2: The number of papers using each imputation method by year

No papers published before 2009 that were included in the study used methods of multiple imputation, and the majority of simple imputation based papers were published in the last three years. This could indicate an increase in a recognition and awareness of the benefits of using imputation within the past few years.

The use of imputation methods based on the percentage of completing patients within the study

With larger percentages of missing data, there is a greater potential for bias if non-completing patients are ignored within the analysis. Figure 3.3 is presented to show the number of patients using each missingness method based on the percentage of completing

patients.

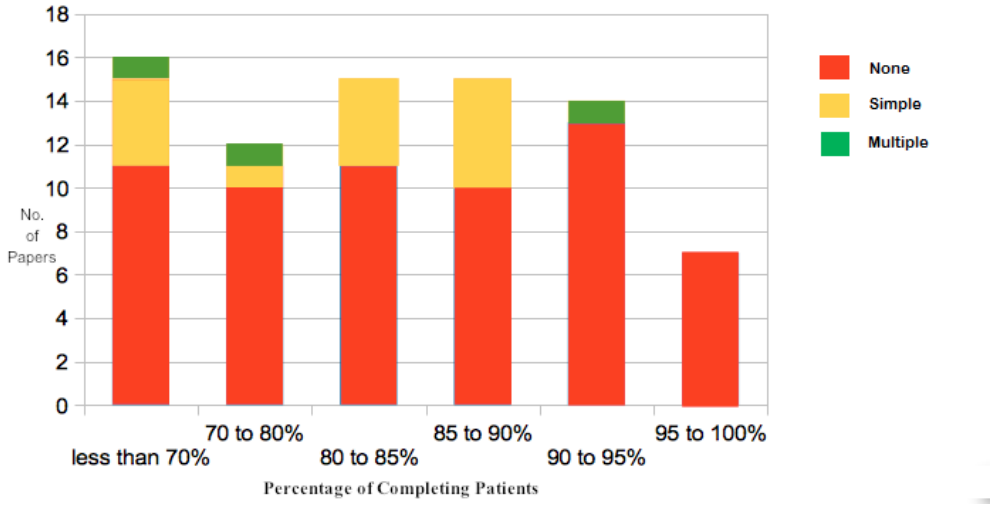


Figure 3.3: The number of papers using each imputation method by percentage of completing patients

Out of the papers with a clear definition of the methods for handling missing data, 12 (12.3%) did not mention the percentage of completing patients. 11(68.8%) papers with less than 70% completing patients used no imputation methods. In general, trials with less than 10% patients dropping out rarely used imputation methods.

Were the reasons for dropout recorded?

The level of information given for dropout in each study was assessed by putting each paper into one of the following categories; "Yes" is defined as detailed discussion on missingness or reason for dropout, including a record of the number of people that dropped out and the specific reasons for dropout recorded at each time point. "Partial Information" is defined as being less detailed with some reference to the reasons for dropout, but not necessarily indicating the number of patients at each time point or providing specific medical reasons. Papers in the "No" category provided no details of the reasons for missing data.

Table 3.2 shows that 35 (39.8%) papers fell in to the "Yes" category, providing substantial and detailed reasons for the missing data of each patient. In 25 (28.4%) cases, the details of reasons for dropout could be categorised as "Partial Information". 28 (31.8%) of papers failed to provide any reasons for patients dropping out. There was not found to be an increase in the quality of missing data reporting for more recent studies with 8 out of 29 papers (27.6%) published in 2011/2012 providing detailed reasons for missing data.

Where no reasons were recorded for missing data, 7 out of the 28 (25.0%) papers made an attempt to explain the reasons for missing data handling method used. For the papers that had some reference to reasons for dropout, 30 out of 60 (50%) made an attempt to explain the reasons for their chosen method of missing data handling. However, the majority of these papers provided a justification of the general statistical benefits of the methods they applied, while few referenced the specific prognostic factors and potential dropout reasons unique to the trial in question.

Assessment of the appropriateness of the missing data methods used

The reasons provided for the use of each missing data method within the 100 papers was extracted, and the appropriateness of each method was assessed on a trial by trial basis. Of the 37 papers that attempted to explain the reasons for the chosen missing data method, 19 (51.3%) had provided sufficient detail and justification for their choice. For 5 papers, it was unclear whether the justifications provided were enough for this systematic review when simply observing the published paper, as the appropriateness was dependent on some features of the specifics of the trial dataset. In 13 (35.1%) of the cases it was felt that the justification provided was insufficient for the method used. In many of these cases, the author provided general advantages of the missing data method, but failed to acknowledge the specific details of the trial in question e.g. "certain relatively simple methods can be

appropriate" referring to a complete case analysis when the trial in question has a high percentage of missing outcome data. While this justification may be appropriate in some trials with low percentages of missing data, despite the explanation being vague, it disregards the fact that the trial in question had over 30% missing data.

Table 3.2 shows an assessment of whether the missing data handling methods used in each RCT were appropriate. In 44 (50%) of the articles where the methods were clear, it was established that the method used to handle missing data was appropriate, e.g. LOCF method for missing data being used in a trial where there was a steady state outcome. It was difficult to determine the appropriateness in 14 (15.9%) of the papers, e.g. trials where the amount of, or nature of missing data was not recorded in the paper. In 30 (34.1%) of the papers it was decided that the method used was not appropriate, e.g. a trial with high percentages of missing data that used a complete case analysis.

In all cases where multiple imputation was used, this appeared to an appropriate method. There were 13 (40.6%) which used a complete case analysis where the percentage of missing data was too high (over 10%) to justify. It was concluded that the majority of papers which used mixed models without imputation and simple imputation methods as a form of missing data handling were justified in doing so.

There were some unexpected trends when observing general trial characteristics in relation to the appropriateness of a missing data method. For example, Figure 3.4 demonstrates whether an acceptable method was used based on the number of patients in the trial.

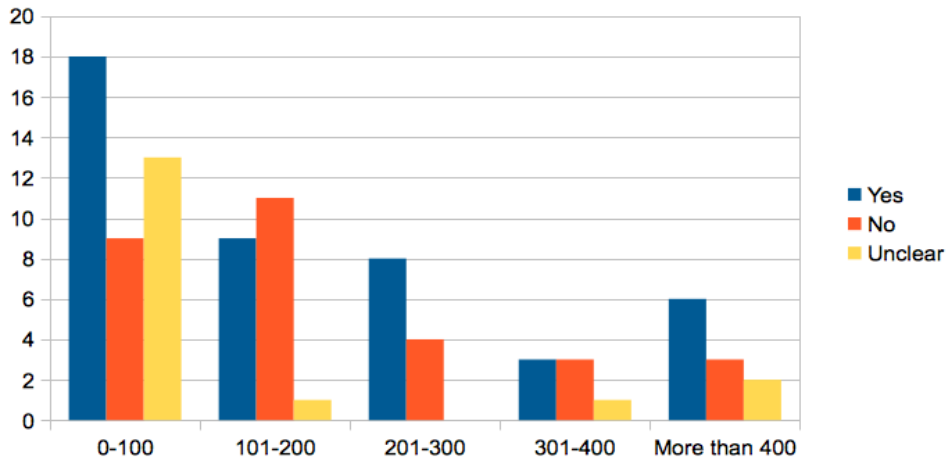


Figure 3.4: The number of papers using appropriate methods by number of patients

Overall, the categories of 0-100, 201-300 and more than 400 patients had the highest proportions of appropriate missing data methods used. There was a disproportionately high percentage of RCTs which used inappropriate missing data handling methods with between 101 and 200 patients randomised. The majority of papers for which it was difficult to assess the quality and appropriateness of missing data handling methods had fewer patients (between 1 and 100), while this was less of an issue for larger trials.

Figure 3.5 shows the relationship between the number of longitudinal time points, and the acceptability of the missing data method.

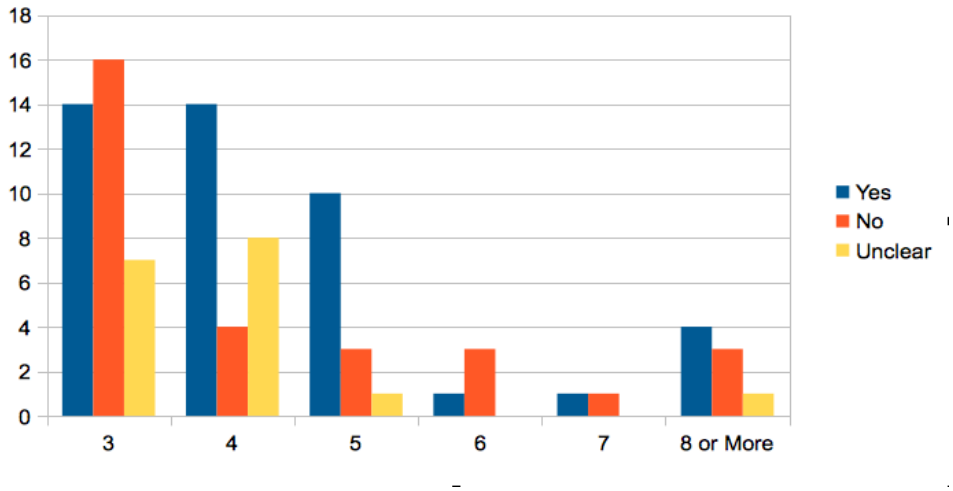


Figure 3.5: The number of papers using appropriate methods by number of longitudinal timepoints

From Figure 3.5, it is observed that a high percentage of inappropriate missing data handling methods were used in studies with only three time points. This may be due to trialists not considering missing data methods for trials where less outcome data is presented. This was less of an issue for the trials with 4 or more time points. The only exception was in trials with 6 time points, although there were few studies in this category.

There appeared to be no differences in preference of missing data handling method or levels of justification for different countries of publication. In all the most popular clinical areas of study within this review there was at least one paper which did not use an appropriate method of missing data handling, and In the area rheumatology, 5 out of 9 (55.6%) of papers used an inappropriate missing data handling methods, as detailed in Table 3.2.

Imputed datasets as a comparison to non-imputed data

Out of the 18 papers that used imputation methods, 11 (61.1%) made no reference to a comparison between the complete case analysis results and the imputed dataset results.

Of the 7 papers that made the comparison, 2 (28.6%) of them yielded different clinical conclusions about treatment effect when complete case analysis was compared to one of the imputed datasets, although one paper only provided the details of this difference for the purpose of illustration. One of the 2 papers (mental health) had 43.2% dropout, and reported the p-values for treatment comparison using the complete case (p=0.428), mean imputation (p=0.360), LOCF (p=0.026), and multiple imputation (p=0.426). The treatment effect was significant when LOCF was used but not when other imputation methods were used. However, in this case the authors suggested that LOCF was not an appropriate method to use for missing outcome data in their trial. The authors suggested that LOCF exaggerated the difference between the treatment groups over time, due to the "non-random pattern of missing values" [98]. Therefore no difference in treatment effect was concluded. The second study, also in mental health, stated that there had been a difference in conclusion without presenting both sets of results. Both of these papers had less than 70% of patients completing the study.

3.4 Discussion

3.4.1 The Extent of Missing Data Handling and Use Of Imputation Methods

In the CONSORT statement, point 13b. states "for each group, losses and exclusions after randomisation, together with reasons" should be included within the trial report [97]. Details of this nature were not presented in a large proportion of articles within the study. It is difficult to suggest a gold standard for missing data handling as the appropriateness of a method is dependent on the unique nature of missing data within each individual trial. However, by carrying out a complete case analysis or eliminating certain patients based on level of missingness assumes that the data is missing completely at random. This is rarely, if ever the case in practice [4]. Therefore it was disappointing to see that 39 out

of 88 (44.3%) RCT analyses excluded certain patients from the analysis. As 32 out of 88 (36.4%) papers used a complete case analysis, this represents a decrease in the use of this method since the study by Wood, White and Thompson in 2004 [89]. Greater understanding of the benefits and methods of multiple imputation have been developed in recent years. However, only 4 papers within the study used multiple imputation, with the evidence suggesting that statisticians are more frequently using simple imputation methods. One positive sign is that the trends suggest that there are a greater number of papers in the past few years using multiple imputation methods. It was also interesting to discover that having a statistician involved within a trial investigation didn't appear to have an impact on the choice of imputation method, or the level of justification for the chosen method.

One factor in determining the optimal missing data handling method in a trial is the percentage of missing data. The systematic review published by Wood, White and Thompson [89] found that imputation was more frequently used in papers with larger amounts of missing data. The updated systematic review in this chapter shows that while papers with a lower percentage of dropout appeared to be less frequent in their use of imputation, there was a large amount of papers with high levels of dropout that did not use imputation. In particular, 11 out of the 16 (68.8%) papers with more than 30% used no imputation methods. Additionally, 30 (33.0%) of papers used inappropriate missing data handling methods, and 13 (35.1%) of the papers that attempted to explain the reasons behind their choice of method provided an inadequate justification.

For the benefit of clarity, one suggested technique is to present imputed dataset results alongside the results for just complete case analysis. This was done in 7 out of 18 (38.9%) papers with imputation methods used. Out of these 7 papers, 2 studies yielded different conclusions when some of the imputed results were compared to the complete case; both of these papers had over 30% dropout. One of the two papers in particular used a multitude of imputation methods, and a significant treatment effect was given when LOCF is used.

While LOCF was not selected as an appropriate missing data handling method of analysis in this paper, this illustrates further how an incorrect choice of missing data handling method can influence the results.

The discussions within this paper have been largely directed towards appropriate methods of statistical analysis, with a particular emphasis on the use of imputation. Missing data becomes less of a problem when larger amounts of information can be obtained on the patients that dropped out. When possible in trials with longitudinal measurements, clinicians and trialists should ensure that the trial design considers the potential for missing data arising in the study and aims to take precautions to try reduce the amount of missingness within a study.

3.4.2 Guidelines for Missing Data Handling - The Four Point Plan

As the issue of missing data in general is not addressed in great detail in the CONSORT statement, it may be of use to suggest the following guidelines in order to formalise a procedure of missing data handling. This is with the interest of ensuring we are obtaining accurate prognostic conclusions by not failing to recognise the problems that come with ignoring missing data within a trial.

- Within a trial the reasons for missing data, and more specifically the reasons for dropout, should be reported in detail. This can be defined as each individual patient's reason for dropout being recorded within the study.
- After assessing these reasons for missingness, detailed discussions should be had as to the methods that will be used for missing data handling.
- These methods should then be justified within the report, and their potential limitations described.
- When the final analysis is carried out, any imputed dataset results should be presented alongside the complete case analysis results.

3.4.3 Missing Data Handling in the MAGNETIC Trial

For an example of how missing data could be handled in a trial, the MAGNETIC trial provides a good demonstration [47]. This trial clearly stated its aims, and the reasons for patient dropout was recorded when possible. Once these reasons were assessed, suitable methods were chosen and results were presented for the complete case analysis, as well as using missing data methods. One particular method employed in this trial which was not used in any of the papers included in this systematic review is "joint modelling of longitudinal and time-to-event data." The benefits of this method have been highlighted in Chapter 2.

By following the example of the MAGNETIC trial, and the 4-point plan proposed, it can be ensured that trialists do not allow the missing data to be a catalyst for inaccurate conclusions within randomised control trials with longitudinal readings.

3.4.4 Limitations

Within any systematic review study there may be potential biases which we have done our best to avoid. This potential bias may be increased by randomly selecting a subset of papers, however we felt that this was a more informative method than selecting only papers from high impact journals. The possibility was considered that higher impact journals may publish trials with greater detail of missing data. While little research has been done to confirm or deny this, not putting a restriction on journal eliminated this potential problem. Should more time have been available, it would be informative to assess the missing data mechanisms of all 381 papers that were regarded as eligible. However, we felt that the 100 included may have provided an accurate overview of the extent of the problem of mishandling of missing data.

3.5 Conclusion

This study indicates that a large proportion of papers fail to recognise the issue of missing data, and many gave insufficient information to ensure that an accurate method of missing data handling was used. The majority of papers failed to explain their reasons for the method of missing data handling employed within their trial. As well as this, less than 40% of papers gave detailed reasons for the missingness. Collecting and presenting the reasons for missing data can prove a valuable and important asset when establishing the consistency of trials, as well as drawing accurate conclusions. There was very little consistency in the levels that different trials considered the problems caused by missing data. In general, a greater awareness is needed in order to ensure that clinical investigators can obtain clinically accurate results from the trial in question by making informed choices and using appropriate methods of missing data handling.

Joint modelling of longitudinal and time-to-event data was not used in any of the papers included in this systematic review, despite the benefits highlighted in Section 1.4. Many papers fail to address the issues caused by dropout appropriately so it could be recommended that joint modelling should be employed more often. In the correct circumstances, the amount of "guesswork" required when establishing successful handling of missing data techniques can be reduced by using these models, and details of dropout can also be modelled alongside the longitudinal outcome.

It is possible that the reason no identified papers used joint modelling is due to a lack of awareness or understanding of the model, or that some important trial design properties for joint models are yet to be addressed in detail in published literature. In the next chapter, we focus on the methodological and trial design aspects of joint modelling by investigating sample size and power calculations for the random slope and intercept joint model.

Chapter 4

Sample Size and Power Calculations in Joint Modelling

4.1 Introduction

In Chapters 1 and 2, a methodology for joint models has been presented and the benefits of applying these methods to simultaneously monitor a longitudinal outcome and dropout has been justified for certain trial designs. In the field of joint modelling, the published literature has focused primarily on the development of model specifications [42], [44], [10], while the topic of trial design for joint longitudinal and event time outcomes is rarely discussed. When planning a trial to be analysed using joint modelling methods, the same design considerations should be addressed prior to a study as with any other type of statistical analysis [1]. In particular it is important to generate a sample size which ensures that enough individuals are included in the study to detect a clinically significant difference, but that also minimises the risk of patients being unnecessarily exposed to an experimental treatment [99].

Currently, there are several sample size formulae derived in literature for separate longitudinal and time-to-event data. A summary of the work done in the area of longitudinal data is highlighted in Diggle (2002) [22], while Schoenfeld [100] originally derived a sample size formula for the Cox-Proportional Hazards model in 1983, with the other developments and specifications since presented in Therneau and Grambsch (2000) [101]. However, for models which account for both types of data, greater considerations must be given when deriving and estimating the power and sample size within a study.

While Chen et al (2011) [48] derived a sample size formula for the estimation of overall treatment effect in the general polynomial joint model [48], little work has been done on the development of sample size formulae in joint modelling for other specifications. In particular no sample size formulae or power formulae have been generated for the Henderson et al. [10] random slope and intercept joint model, which is the primary focus of this thesis. When fitting this model, there are generally three main parameters of interest;

the longitudinal treatment effect, β_1, the time-to-event treatment effect, β_2 and the link between the longitudinal and time-to-event outcomes, γ.

The systematic review carried out in Chapter 3 showed that joint modelling of longitudinal and time-to-event data is rarely used in practice. Currently, sample size calculations for this type of modelling are done using simulations [140]. However, this may be one of the reasons that joint modelling is not used more often. Clinicians may be unaware of how to conduct these pre-trial simulations or be put off by what they perceive to be complicated statistical programming.

For the Henderson et al. specification of the joint model [10] , no research has focused on how different parameters and trial properties affect the power for β_1, β_2 and γ. In this chapter, sample size formulae for γ and β_2 are derived for the aforementioned random slope and intercept joint model using the distribution of the Rao score statistic [102]. Furthermore, a discussion is invoked about the potential factors affecting the sample size for β_1.

Using a simulation study based on the parameters of MAGNETIC, the success of the newly generated sample size and power formulae are tested, and the properties of the power for β_1 investigated. As a final task, the power for each parameter in the MAGNETIC trial will be calculated.

4.2 Sample Size and Power for β_2

Within the specification in Section 1.4.2, the time-to-dropout is modelled using a Cox-Proportional Hazards model, in which β_2 is defined as the estimate of the the log-hazard ratio between treatments. Schoenfeld derived a sample size formula for β_2 in a standard Cox-Proportional Hazards model, which does not model interactions with a longitudinal outcome monitored over time, and contains no patient specific random effects [100]. To

derive a sample size formula for β_2 in the joint modelling context, we apply similar methods to those used by Schoenfeld, by using the Rao score test to establish the distribution of the Rao test statistic.

The longitudinal element of the random slope and intercept model can be written as

$$Y_i(t) = X_{1i}(t)\beta_1 + W_{1i}(t) \tag{4.1}$$

where $W_{1i}(t) = U_{1i} + U_{2i}t$. The time-to-dropout is modelled by

$$\lambda_i(t) = \lambda_0(t)\exp\{X_{2i}(t)\beta_2 + W_{2i}(t)\}. \tag{4.2}$$

The link between the latent parameters defined as $W_{2i}(t_{ij}) = \gamma W_{1i}(t_{ij})$. We begin by making the assumption that U_1 and U_2 have no impact on the power for β_2, and address the interactions with random effects later in the chapter. While $X_{2i}(t)$ can generally represent the individual patient values for a collection of variables, for the derivation of sample size we can define $X_{2i}(t)$ as a binary indicator indicating the treatment group and β_2 is the corresponding treatment effect as Schoenfeld demonstrated that the other independent fixed variables will have no effect on the power for β_2 [100]. Therefore, for the derivation let $X_{2i}(t)$ and $X_{1i}(t)$ be the same (the treatment indicator), so we denote this as $X_i(t)$. As the treatment is the same at all time points for each patient i, we can denote this as X_i. Therefore rewrite (4.2) as;

$$\lambda_i(t) = \lambda_0(t)\exp\{X_i\beta_2 + \gamma(U_{1i} + U_{2i}t)\} \tag{4.3}$$

We assume a two arm trial and label the treatments either A or B. Define P_A to be the probability that patient i is in treatment group A. In general, the score function is defined by:

$$\mathscr{U}(\theta) = \frac{dlogL(\theta|x)}{d\theta} \tag{4.4}$$

for the parameter of interest θ (in this case β_2) and data x. The Fisher Information is given by

$$I(\theta) = -E[\frac{d^2}{d\theta^2} logL(X;\theta)|\theta] \tag{4.5}$$

where L denotes the corresponding likelihood. The statistic to test $H_0 : \theta = \theta_0$ is then

$$S(\theta_0) = \frac{\mathscr{U}(\theta_0^2)}{I(\theta_0)}. \tag{4.6}$$

This is asymptotically normal when H_0 is true. Rao [103] also showed that this was mathematically equivalent to testing $S^*(\theta) = \sqrt{S(\theta)}$, which follows a normal distribution. The power and number of events required for β_2 in a trial which uses joint models can be established by finding the distribution of $S(\beta_2)$. The partial likelihood for Equation (4.3) is

$$L_i = \left\{ \frac{\exp(\beta_2 X_k + \gamma(U_{1i} + U_{2i}t))}{\sum_{k=1}^{N} I(S_k \leq S_i) \exp(\beta_2 X_k + \gamma(U_{1i} + U_{2i}t))} \right\}^{\Delta_i} \tag{4.7}$$

where N is the total number of patients in the trial, S_i defines the event-time for patient i and Δ_i is the indicator function which is equal to 1 if an event occurred and 0 otherwise [100].

From Schoenfeld and Chen [48, 100], we can define G_i for any function $g(X, Y)$ as;

$$G_i\{g(X,Y)\} = \frac{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\gamma(U_{1i} + U_{2i}t)) g(X,Y)}{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\gamma(U_{1i} + U_{2i}t))} \tag{4.8}$$

which is related to the partial likelihood with $\beta_2 = 0$. It has been shown for Cox proportional hazards models [100] that the score function can be written as

$$S(\beta_2) = \frac{N^{-\frac{1}{2}} \sum_{i \in D} \{X_i - G_i(X_k)\}}{\{N^{-1} \sum_{i \in D} G_i\{X_k^2\} - (G_i\{X_k\})^2\}^{\frac{1}{2}}} \tag{4.9}$$

where $i \in D$ denotes the set of subjects who experienced an event, and $k = 1 \ldots N$.

Now we define

$$e_i(X_k) = \frac{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\beta_2 X_k + \gamma(U_{1i} + U_{2i}t))X_k}{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\beta_2 X_k + \gamma(U_{1i} + U_{2i}t))} \tag{4.10}$$

This can be used to rewrite $S(\beta_2)$ as

$$S(\beta_2) = \frac{N^{-\frac{1}{2}} \sum_{i \in D} \{X_i - e_i(X_k)\}}{\{N^{-1} \sum_{i \in D} G_i\{X_k^2\} - (G_i\{X_k\})^2\}^{\frac{1}{2}}} \tag{4.11}$$

$$+ \frac{N^{-\frac{1}{2}} \sum_{i \in D} \{e_i(X_k) - G_i(X_k)\}}{\{N^{-1} \sum_{i \in D} G_i\{X_k^2\} - (G_i\{X_k\})^2\}^{\frac{1}{2}}}$$

$N^{-\frac{1}{2}} \sum_{i \in D} \{X_i - e_i(X_k)\}$ is asymptotically normal with mean 0 as it is the score function of the partial likelihood, with variance equal to $N^{-1} \sum_{i \in D} e_i\{X_k^2\} - (e_i\{X_k\}^2)$ [48]. As in Schoenfeld, we can treat β_2 as $O(n^{-\frac{1}{2}})$ [100]. Therefore $e_i\{(X_k^2)\} \to G_i\{(X_k^2)\}$ when $\beta_2 \to 0$, so therefore the first term of (4.11) tends to N(0,1) as $\beta_2 \to 0$ [100]. Now focusing on the second part of the numerator $N^{-\frac{1}{2}} \sum_{i \in D} \{e_i(X_k) - G_i(X_k)\}$, using a Taylor series expansion about $\beta_2 = 0$ to the second order, it can be observed that the numerator approximates to

$$e_i\{X_k\} - G_i(X_k) \approx \beta_2 \{G_i\{X_k^2\} - (G_i\{X_k\})^2\}. \tag{4.12}$$

By substituting this into the second term, and simplifying with the denominator, it is found that the second term approaches

$$\beta_2 \left\{ \sum_{i \in D} \{G_i\{X_k^2\} - (G_i\{X_k\})^2 \right\}^{\frac{1}{2}} \tag{4.13}$$

Initially, we derive a sample size formula based on the assumption that $Var(U_2)$ has no impact on the power, and adjust for this impact post-derivation. Therefore when $\beta_2 \to 0$, we find that $G_i\{X_k\} \to E\{X_k\}$, and therefore $G_i\{X_k^2\} - (G_i\{X_k\})^2 \to Var(X_k) = P_A(1 -$

P_A), where P_A is the proportion of patients randomised to treatment group A. Hence $S(\beta_2)$ is asymptotically normal with a mean

$$\beta_2 D^{\frac{1}{2}} \left\{ \frac{1}{D} \sum_{i \in D} P_A(1 - P_A) \right\}^{\frac{1}{2}} = \beta_2 (D P_A (1 - P_A))^{\frac{1}{2}} \tag{4.14}$$

and a variance 1 where D is the number of events.

Therefore [105], for a two sided test with significance α and power β,

$$\frac{\beta_2 (D P_A (1 - P_A))^{\frac{1}{2}}}{1} \approx z_\beta + z_{1-\alpha/2}$$

and hence the number of events required for a two sided test with significance α and power β is given by:

$$D = \frac{(z_\beta + z_{1-\alpha/2})^2}{P_A(1 - P_A){\beta_2}^2}. \tag{4.15}$$

This derivation makes the assumption that the parameters of U_1 and U_2 have no impact on the sample size formula. It has been shown that independent and fixed non time-parameters have no influence on this sample size, so by treating U_1 in this way we can rule out the sample size formula being affected by U_1 values [104]. However this may not be the case for $Var(U_2)$ as this introduced a time influence into the Cox-Proportional Hazards element of the model. Therefore, the simulation study in Section 4.6 investigates the effect of varying $Var(U_2)$ on the number of required events for β_2. Investigations of the power for a Cox-Proportional Hazards model which include a patient specific random slope component has not been carried out in published literature.

4.3 Sample Size and Power for γ

In some cases, a trial may be focused on establishing a relationship between a given biomarker and dropout [40], which can be attained by estimating γ. More details of the nature and properties of γ will be provided in Chapter 5. So far, there have been no derivations for the number of events required to successfully estimate the γ parameter for joint models.

In this section, a sample size formula is derived for the γ parameter of the Henderson et al. [10] specification. To do so, the score statistic for the Rao score test is derived and the distribution is established. The time-to-dropout element of the model is defined as in Equation (4.3). For the reasons highlighted in the β_2 derivation, we can take β_1 and $\beta 2$ to be the parameters for the treatment effects in the longitudinal and time-to-dropout model. Therefore, substituting (4.1) into (4.3) we can write:

$$\lambda_i(t) = \lambda_0(t) \exp\{X_i(t)'\beta_2 + \gamma(Y_i - \beta_1 X_{1i})\}. \tag{4.16}$$

As in the previous derivation, we can define a formulae for $G_i(g(X,Y))$ and $e_i(g(X,Y))$. In this case, in conjunction the derivation of Equation (4.15), define

$$G_i\{Y_k(S_i) - \beta_1 X_k\} = \frac{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\beta_2 X_k)(Y_k(S_i) - \beta_1 X_k)}{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\beta_2 X_k)} \tag{4.17}$$

and

$$e_i\{Y_k(S_i) - \beta_1 X_k\} = \frac{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\gamma(Y_k(S_i) - \beta_1 X_k) + (\beta_2 X_k))(Y_k(S_i) - \beta_1 X_k)}{\sum_{k=1}^{N} I(S_k \geq S_i) \exp(\gamma(Y_k(S_i) - \beta_1 X_k) + (\beta_2 X_k))} \tag{4.18}$$

Therefore, the partial likelihood is derived as

$$L_i = \left\{ \frac{\exp\{\gamma(Y_i - \beta_1 X_i) + \beta_2 X_i\}}{\sum_{k=1}^{N} I(S_k \leq S_i) \exp\{\gamma(Y_i - \beta_1 X_i) + \beta_2 X_i\}} \right\}^{\Delta_i} \tag{4.19}$$

For the simplicity, we define $Q_k(S_i) = Y_k(S_i) - \beta_1 X_k = U_{1k} + U_{2k} S_i$. Therefore the score statistic for the Cox's partial likelihood is given by

$$S(\gamma) = \frac{N^{-\frac{1}{2}} \sum_{i \in D} \{Q_i(S_i) - G_i\{Q_k(S_i)\}\}}{\{N^{-1} \sum_{i \in D} (G_i\{Q_k(S_i)^2\} - G_i\{Q_k(S_i)\}^2)\}^{\frac{1}{2}}} \tag{4.20}$$

The score statistic $S(\gamma)$ can be written as

$$S(\gamma) = \frac{N^{-\frac{1}{2}} \sum_{i \in D} \{Q_i(S_i) - e_i\{Q_k(S_i)\}\}}{\{N^{-1} \sum_{i \in D} (G_i\{Q_k(S_i)^2\} - G_i\{Q_k(S_i)\}^2)\}^{\frac{1}{2}}} \tag{4.21}$$

$$+ \frac{N^{-\frac{1}{2}} \sum_{i \in D} \{e_i\{Q_k(S_i)\} - G_i\{Q_k(S_i)\}\}}{\{N^{-1} \sum_{i \in D} (G_i\{Q_k(S_i)^2\} - G_i\{Q_k(S_i)\}^2)\}^{\frac{1}{2}}}$$

$\sum_{i \in D} \{Q_i(S_i) - e_j\{Q_k(S_i)\}$ is the score function of the partial likelihood L_i defined in Equation (4.19). Consequently the the numerator of the first row of (4.21) is asymptotically normal with mean 0 and a variance equal to $N^{-1} \sum_{i \in D} [e_i\{Q_k(S_i)^2\} - (e_i\{Q_k(S_i)\})^2]$. As $\gamma \to 0$, $e_i\{(Q_k(S_i))^q\} \to G_i\{(Q_k(S_i))^q\}$ where q is any integer and therefore this first term $\to N(0,1)$. We expand the numerator of the second term in a Taylor series about $\gamma = 0$. This demonstrates that

$$e_j\{Q_k(S_i)\} - G_i\{Q_k(S_i)\} \approx \gamma\{G_i\{Q_k(S_i)^2\} - (G_i\{Q_k(S_i)\})^2\} \tag{4.22}$$

so the second term approaches

$$\gamma \left\{ \sum_{i=1}^{D} G_i\{Q_k(S_i)^2\} - (G_i\{Q_k(S_i)\})^2 \right\}^{\frac{1}{2}}. \tag{4.23}$$

If it is assumed that each treatment group is large, then

$$G_i\{(Q_k(S_i))^q\} = \frac{\frac{1}{N}\sum_{k=1}^{N} I(S_k \geq S_i)\exp(X_k\beta_2)(Q_k(S_i))^q}{\frac{1}{N}\sum_{k=1}^{N} I(S_k \geq S_i)\exp(X_k\beta_2)} \tag{4.24}$$

$$\rightarrow \frac{E\{I(S_k \geq S_i)(Q_k(S_i))^q\}}{E\{I(S_k \geq S_i)\}} \tag{4.25}$$

S_k is independent of U_1 and U_2 when $\gamma = 0$ and $I(S_k \geq S_i)$ is independent of $Q_k(S_i)$ conditional on S_i. Therefore

$$G_i\{(Q_k(S_i))^q\} \rightarrow E_i\{(Q_k(S_i))^q\} \tag{4.26}$$

In conjunction with Equation (4.23), as $\gamma \rightarrow 0$:

$$\{G_i\{Q_k(S_i)^2\} - (G_i\{Q_k(S_i)\})^2\} \rightarrow Var\{Q_k(S_i)\} \tag{4.27}$$

Hence;

$$\gamma D^{\frac{1}{2}} \left\{ \frac{1}{D}\sum_{i\in D}\{G_i\{(Q_k(S_i))^2\} - (G_i\{Q_k(S_i)\})^2 \right\}^{\frac{1}{2}} \tag{4.28}$$

$$\rightarrow \gamma D^{\frac{1}{2}} \left\{ \frac{1}{D}\sum_{i\in D} Var(Q_k(S_i)) \right\} \tag{4.29}$$

The variance of $Q_k(S_i)$ can be calculated as

$$Var(Q_k(S_i)) = Var(U_{1k} + U_{2k}S_i) \tag{4.30}$$

$$= Var(U_{1k}) + Var(U_{2k})E\{I(t \leq t_f)T^2\}/\tau + 2Cov(U_{1k}, U_{2k})E\{I(t \leq t_f)T\}/\tau$$

where T is the event or censoring time, $\tau = D/N$ and $E\{I(t \leq t_f)T\}$,$E\{I(t \leq t_f)T^2\}$ are truncated moments.

Therefore $S(\gamma)$ is asymptotically normal with unit variance and the mean is equal to $\gamma\{D\sigma_s^2\}^{\frac{1}{2}}$. $D \to \infty$. Therefore [105], for a two sided test with significance α and power β,

$$\frac{\gamma\{D\sigma_s^2\}^{\frac{1}{2}}}{1} \approx z_\beta + z_{1-\alpha/2}$$

and hence the number of events required for a two sided test with significance α and power β is given by:

$$D = \frac{(z_\beta + z_{1-\alpha/2})^2}{\sigma_s^2 \gamma^2} \tag{4.31}$$

where $\sigma_s^2 = Var(U_{1k}) + Var(U_{2k})E\{I(t \leq t_f)T^2\}/\tau + 2Cov(U_{1k}, U_{2k})E\{I(t \leq t_f)T\}/\tau$.

Due to the multiplicative nature of γ and $\{Var(U_{1k}),\ Var(U_{2k})\}$ in the model, the values of $Var(U_{1k})$, $Var(U_{2k})$ and $Cov(U_{1k}, U_{2k})$ are required to generate the sample size. If these values are unknown, then Bayesian techniques as described by Tsiatis et al (1995) can be used to estimate these variances and the potential sample size [40,48]. When U_1 and U_2 are independent, then in Equation (4.31) $\sigma_s^2 = Var(U_{1k}) + Var(U_{2k})E\{I(t \leq t_f)T^2\}/\tau$.

Estimates of the truncated moments $E\{I(t \leq t_f)T\}$ and $E\{I(t \leq t_f)T^2\}$ are also required to calculate sample size. In some trials with time-to-event or survival outcomes, it is possible to obtain an estimate for these based upon previous literature. However, when dropout is the event of interest estimates of the truncated moments can rarely be obtained from previous studies, particularly when a new treatment is being tested in an RCT.

Therefore, we propose approximating $E\{I(t \leq t_f)T\}$ and $E\{I(t \leq t_f)T^2\}$, by assuming a discrete uniform distribution of T for the dropouts, such that patients are equally likely to drop out at any point in the study. While this may not be the case in practice, and time points will not always be equally spaced, this is sufficient for generating an approximate sample size formula and investigating the impact of more time points on the power calculations.

Firstly we derive an approximation for $E\{T_{drops}\}$ for patients that dropped out, based on the expectation of the generalised discrete uniform distribution [106], which is calculated as the average of the first and last dropout point (0 and $t_f - \frac{t_f}{ntms-1}$), where $ntms$ is the number of time points.

$$\begin{aligned} E\{T_{drops}\} &= \frac{0 + t_f - \frac{t_f}{ntms-1}}{2} \\ &= \frac{t_f(1 - \frac{1}{ntms-1})}{2} \end{aligned}$$

As this is the expectation across the dropouts only, the expectation $E\{I(t \leq t_f)T\}$ would be

$$E\{I(t \leq t_f)T\} = \tau \left\{ \frac{t_f(1 - \frac{1}{ntms-1})}{2} \right\} \tag{4.32}$$

The second truncated moment is calculated using $E\{T^2_{drops}\} = E(T_{drops})^2 + Var(T_{drops})$ from the definition of variance. For simplicity, let $d = \frac{t_f}{ntms-1}$ denote the distance between the longitudinal timepoints in the study. Then, by the definition of the expectation and variance of the generalised discrete uniform distribution:

$$
\begin{aligned}
E\{T^2_{drops}\} &= E(T_{drops})^2 + Var(T_{drops}) \\
&= \left\{\frac{t_f - d}{2}\right\}^2 + d^2\frac{(ntms-1)^2 - 1}{12} \\
&= \left\{\frac{t_f - d}{2}\right\}^2 + d^2\left(\frac{(ntms-1)^2}{12} - \frac{1}{12}\right) \\
&= \left\{\frac{t_f - d}{2}\right\}^2 + d^2\left(\frac{(t_f)^2}{12}d^{-2} - \frac{1}{12}\right) \\
&= \frac{1}{4}(t_f - d)^2 + \frac{1}{12}(t_f^2 - d^2) \\
&= \frac{1}{4}((t_f)^2 - 2t_f d + d^2) + \frac{1}{12}((t_f)^2 - d^2) \\
&= d^2\left(\frac{1}{4} - \frac{1}{12}\right) - \frac{t_f}{2}d + t_f^2\left(\frac{1}{12} + \frac{1}{4}\right) \\
&= \frac{1}{6}d^2 - \frac{t_f}{2}d + t_f^2\left(\frac{1}{3}\right)
\end{aligned}
$$

By substituting $d = \frac{t_f}{ntms-1}$ back into the equation:

$$
\begin{aligned}
E\{T^2_{drops}\} &= \frac{1}{6}\left(\frac{t_f}{ntms-1}\right) - \frac{t_f}{2}\left(\frac{t_f}{ntms-1}\right) + (t_f)^2\left(\frac{1}{3}\right) \\
&= (t_f)^2\left(\frac{1}{6}\left(\frac{1}{ntms-1}\right)^2 - \frac{1}{2}\left(\frac{1}{ntms-1}\right) + \frac{1}{3}\right)
\end{aligned}
$$

and therefore

$$
E\{I(t \le t_f)T^2\} = \tau\left\{(t_f)^2\left\{\frac{1}{6}\left(\frac{1}{ntms-1}\right)^2 - \frac{1}{2}\left(\frac{1}{ntms-1}\right) + \frac{1}{3}\right\}\right\} \tag{4.33}
$$

using the same technique as in Equation (4.32).

Therefore, sample size for γ can be approximated by

$$\hat{D} = \frac{(z_\beta + z_{1-\alpha/2})^2}{\sigma_s^2 \gamma^2} \tag{4.34}$$

where

$$\sigma_s^2 = Var(U_{1k}) + Var(U_{2k}) \left\{ (t_f)^2 \left(\frac{1}{6} \left(\frac{1}{ntms - 1} \right)^2 - \frac{1}{2} \left(\frac{1}{ntms - 1} \right) + \frac{1}{3} \right) \right\}$$
$$+ 2Cov(U_{1k}, U_{2k}) \left\{ t_f \left(\frac{\left(1 - \frac{1}{ntms-1}\right)}{2} \right) \right\}$$

To verify and investigate formulae (4.31) and (4.34), a simulation study is carried out in Section 4.6 to test both the derived formula based on the data, and the approximation.

4.4 Sample Size and Power for β_1

In many clinical trials with longitudinal data measured, the outcome of interest will be the treatment effect over time within a repeated measures setting [107–109]. However, estimating the power or sample size for a longitudinal treatment effect over time is generally reliant on some assumptions that are untestable prior to the trial commencing. Many different sample size formulae have been proposed for longitudinal data alone [22,110,111]. In the previous chapter, it was shown that the most common method for analysis in longitudinal studies is to analyse the complete cases [60, 89]. In these trials, the approach to sample size is to calculate the number of patients required in total and then increase the number of patients admitted by a figure relating to the expected percentage of dropout. However, even in simple longitudinal trials, assumptions must be made about the within patient correlation in order to determine sample size [22].

Further complications arise when attempts are made to calculate sample sizes for trials that employ more sophisticated methods than a complete case analysis. The wide range

of possible causes of missing data presents a challenge for alternative longitudinal designs as the pattern of missingness can have an effect on sample size [4]. Patients dropping out, staggered or intermittent missing data and missed hospital visits are among many of the potential reasons for incomplete data sets and making predictions about the nature of missing data can be difficult prior to a trial commencing. Approaches for assessing sample size and power calculations for longitudinal trials with missing data have been investigated in published literature using Bayesian methods, frequentist based algorithmic concepts and simulation studies for testing longitudinal treatment effect [48].

When considering a derivation of sample size for β_1 in the joint modelling context, there are a number of variables to be taken into account. Due to the complexities, Rizopoulos (2010) states that simulations should be used to calculate the required sample size [140]. However, this has not been done in practice for an RCT using the joint model defined in Section 1.4.3. Also no investigations have been made into the factors which affect the sample size of β_1 in the context of joint modelling, which is one of the aims of this chapter.

As a preliminary, Diggle et. al. 2002 [22] derived the following sample size formula for standard linear mixed models:

$$N = \frac{2(z_{1-\alpha/2} - z_{\beta})^2(1 + (ntms - 1)\rho)}{\left(\frac{\beta_1}{\sigma}\right)^2 \times ntms} \tag{4.35}$$

where $ntms$ is the number of time points, ρ is the assumed correlation between the repeated measures and σ^2 is the common variance between the two groups. Observing the above formula it can be noted that in this setting, increasing the common variance and assumed correlation between repeated measures will result in a decrease of power. Also as $\rho < 1$, a higher number of time points would result in a higher power for the same number of patients, due to more information being available about the outcome variable.

As highlighted above, more complex derivations have been used to handle missing data when greater detail is available about the nature of the missingness, however sample sizes for longitudinal designs are reliant on information that may not be available prior to a trial commencing. Equation (4.35) serves to provide a basic motivation to establish the effects upon power for β_1 in this study. In Section 4.5, a simulation study is conducted to establish the effect of varying different parameters and trial properties on the power of β_1.

4.5 Simulation Study

A power and sample size simulation study is carried out based on the parameters of the joint model as specified in Section 1.4.3. The main aims of this study can be categorised as

- (a) Testing the accuracy of the previously derived sample size formulae.
- (b) Establishing an approximate relationship between power of β_2 and $Var(U_1)$
- (c) Determining the factors that effect the power for the longitudinal treatment effect β_1.
- (d) Establishing the effect of the different number of time-points on the power for each parameter.
- (e) Estimating the power for each parameter of the MAGNETIC data when using joint modelling.

Simulations will be generated for various different trial properties and parameter values, and the corresponding powers for β_1, β_2 and γ will be calculated.

4.5.1 Methods

Data is simulated based on the parameters of the MAGNETIC dataset in the joint modelling framework using the same simulation methods as in Chapter 2.4.1, to mimic a two-arm trial. Continuous longitudinal data is simulated from Equation (1.1) in Chapter 1,

jointly with time-to-dropout using a flat baseline hazard and a random slope and intercept model from Equation (1.2). Dropout was modified to be the last previous time-point prior to the simulated event times. For these simulations, parameters are set to $\beta_1 = 0.2048$, $\beta_2 = -0.5$ to mirror the MAGNETIC trial, and dropout is set at 20% by varying the baseline hazard.

Different datasets are simulated by varying the number of patients, the number of longitudinal time points, the values of γ and the variances of U_1 and U_2. The power is defined as the proportion of these simulated datasets that yielded a significant result at the 95% confidence level. Data is generated for n=200, 400, 600, 800 patients total in a trial with an equal number of simulated patients randomised to each arm. The number of timepoints in the study is set to $ntms = 4, 6, 9$ and $\gamma = 0.25, 0.5, 0.75, 1$ for different simulations. The follow up time, t_f, was set to 240, and longitudinal time points were generated by dissecting the interval of $[0, 240]$, $ntms$ times.

Simulated data was generated for 3 different combinations of $(Var(U_1), Var(U_2))$, which are set to $(1.19, 0.0003), (1.19, 0.00003)$ and $(0.6, 0.00003)$. The second of these is approximately the parameters estimated in MAGNETIC, with the other two varying $Var(U_1)$ and $Var(U_2)$ respectively. For the analysis where $Var(U_1, U_2) = (1.19, 0.00003)$, we let $Cov(U_1, U_2) = -0.001$, which was the approximate covariance between the random effects in the MAGNETIC analysis. In the other analyses, the covariance was set to be the equivalent of this value for the given variances such that the correlation between U_1 and U_2 was the same for all simulations. Therefore $Cov(U_1, U_2) = -0.0032$ for the set of simulations with $Var(U_1, U_2) = (1.19, 0.0003)$ and $Cov(U_1, U_2) = -0.0007$ for $Var(U_1, U_2) = (0.6, 0.00003)$. The error variance was set to 0.5 for all simulations. For each combination of trial properties, simulated data was analysed using the random slope and intercept joint model, with confidence intervals being calculated using bootstrapping methods with 1000 iterations for β_1, β_2 and γ. Simulations were done using the CONDOR

computing service for R in order to minimise amount of time required.

For β_2, the empirical powers are calculated from the simulations by observing the proportion of significant parameter estimates for different values of $Var(U_2)$ and γ. These are compared to the expected powers estimated by Equation (4.15). This is with a view to testing the validity of the formula, and establishing the impact of varying the random effects and γ on the power of β_2.

For γ the empirical power calculated is calculated from the simulations in the same way, and these are compared to the estimated power from Equations (4.31) and (4.34). The aim is to test the accuracy of the derived formulae based on the both the uniform approximation to the truncated moments (4.34) and the mean empirical estimate of the moments (4.31). We will also establish the effects of varying $Var(U_1), Var(U_2)$ on the power for γ.

For β_1, the empirical power for different simulation settings are generated, and these are compared to the power generated by using a complete case analysis; the most common method for handling missing data according to the systematic review in Chapter 3. This is to establish the extent to which using joint models improves the power when estimating β_1.

4.5.2 Results

β_2 Results

Data was simulated, and the powers of β_2 were calculated for three different combinations of $Var(U)$. The results presented in Tables 4.2 - 4.5 show that as the variances of U_2 were increased, the power decreased. Varying $Var(U_1)$ seemed to have a negligible effect on power, which coincides with the hypothesis given in Section 4.2. It can also be observed from the tables that the number of time points in a study had no impact on the power for β_2.

In general, the formula for power given in Equation (4.15), which relies on the assumption that $\gamma = 0$, predicted a higher power than was given in the simulations. Similarly, the power appeared to be inversely proportional to the value of γ, as when γ was increased, the power of the trial decreased in the majority of cases. To investigate this relationship further, the mean power over the three different numbers of time points was calculated for each combination of γ and $Var(U_1), Var(U_2)$, as $ntms$ was found to have no effect on power. Figure 4.1 shows the relationship between γ and power for β_2 in each simulation when combining the powers for the different time points. The dotted line in the plots indicates the power as predicted by Equation (4.15).

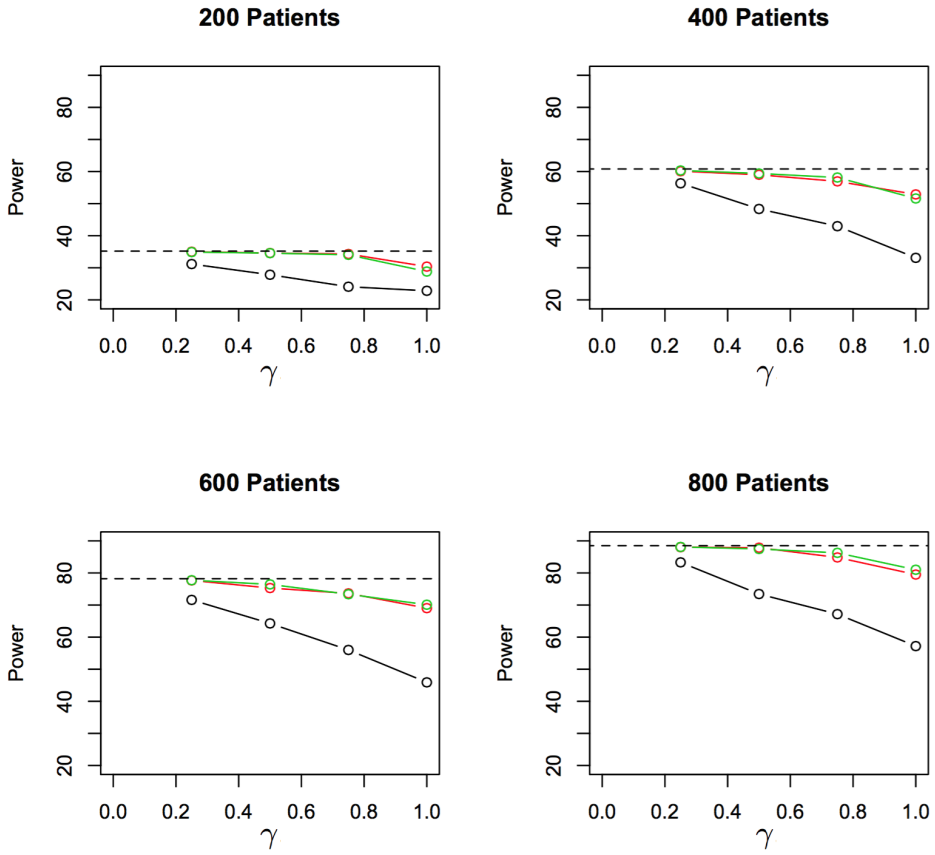


Figure 4.1: Plot of β_2 power against γ by $Var(U)$ combination. $Var(U_1, U_2) = (1.19, 0.0003)$(black), $Var(U_1, U_2) = (1.19, 0.00003)$(red), $Var(U_1, U_2) = (0.6, 0.00003)$(green)

In Figure 4.1 it can be observed that there is a greater loss in power, when compared to the derived formula in Equation (4.15), for the simulations with a higher value of $Var(U_2)$, while the simulations in which $Var(U_1)$ are varied yield similar results to each other. For higher values of γ, the loss in power is also higher. Furthermore, the reduction in power appears to be greater between $\gamma = 0.75$ and $\gamma = 1$ than for small γ values. Figure 4.2 shows the loss in power when comparing the empirical estimates to Equation (4.15) for each $Var(U_1, U_2)$ combination.

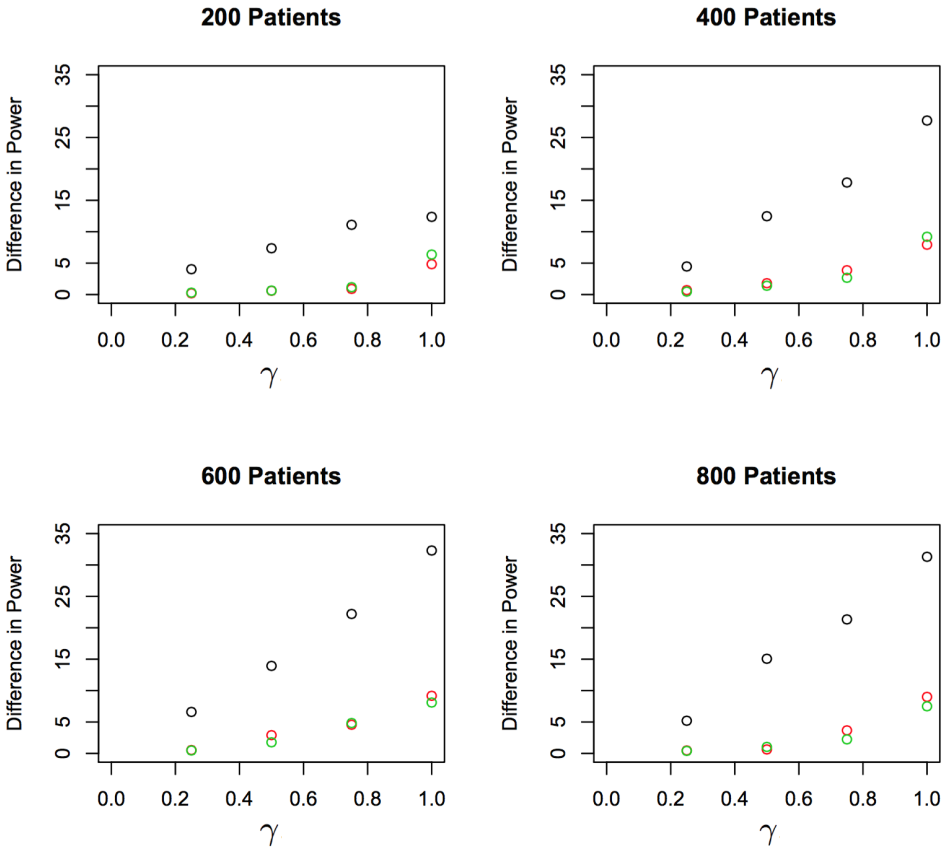


Figure 4.2: Loss in power from predicted power for each $Var(U)$ combination. $Var(U_1, U_2) = (1.19, 0.0003)$(black), $Var(U_1, U_2) = (1.19, 0.00003)$(red), $Var(U_1, U_2) = (0.6, 0.00003)$(green)

Figure 4.2 indicates that when $\gamma = 0$, Equation (4.15) may have given an accurate prediction of the power for β_2. However, the formula gives an overestimation when $\gamma \neq 0$. To compensate for this, Hsieh and Lavori (2000) [104] introduced a variance inflation factor (VIF) to be estimated within a trial for survival analysis with non-independent covariates. This can be estimated as D/D^* where D* is the number of events in a trial, and D is the required number of events to obtain the power of the trial without considering the effect of other parameters on the variance (in this case, from Equation (4.15)). To investigate the relationship between γ, $Var(U_2)$ and power further, the VIF for each simulation was calculated compared to the estimated power. As $Var(U_2)$ and γ are multiplicative within

the trial, $Var(\gamma U_2) = \gamma^2 Var(U_2)$. Figure 4.3 shows plots of the calculated VIF against $\gamma^2 Var(U_2)$.

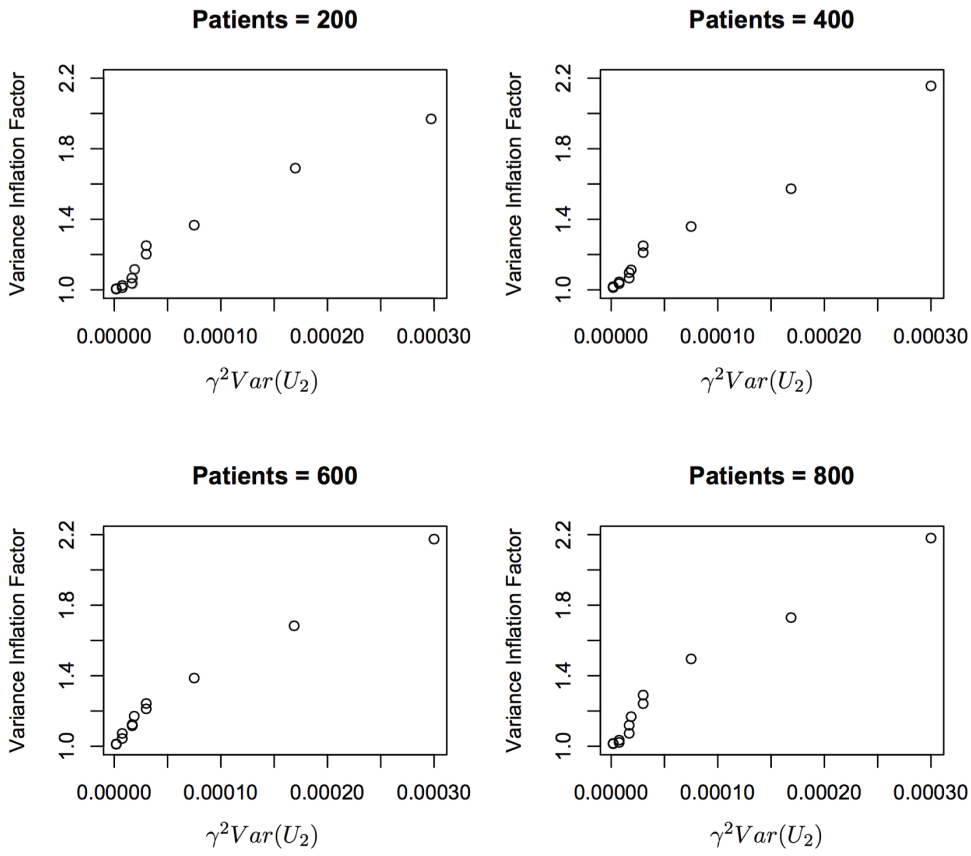


Figure 4.3: Power study for β_2

From Figure 4.3 it can be observed that the relationship between VIF and $\gamma^2 Var(U_2)$ appears to be approximately linear. Therefore, we fit a linear model to the above data to investigate the relationship. Calculating the coefficients for this model to examine the relationship between $\gamma^2 Var(U_2)$ and VIF showed that as $\gamma^2 Var(U_2)$ increased by 0.001, the VIF increases by 3.773. The intercept was found to be $1.00305 \approx 1$ as expected.

From the above simulations it can be deduced that for a follow up time of 240, an approximate formula for the number of events required in Equation (4.15) can be modified

to

$$D = VIF\frac{(z_\beta + z_{1-\alpha})^2}{P_A(1-P_A){\beta_2}^2} \tag{4.36}$$

so

$$D \approx \left(1 + 3773.3\gamma^2 Var(U_2)\right)\frac{(z_\beta + z_{1-\alpha/2})^2}{P_A(1-P_A){\beta_2}^2} \tag{4.37}$$

for a design similar to the MAGNETIC trial.

	$(Var(U_1), Var(U_2)) = (1.19, 0.0003)$			$(Var(U_1), Var(U_2)) = (1.19, 0.00003)$			$(Var(U_1), Var(U_2)) = (0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	31.0	32.0	30.5	34.1	35.0	35.9	34.2	35.4	35.1
γ=0.5	28.1	27.6	27.8	34.6	34.2	35.0	35.2	34.0	34.5
γ=0.75	23.4	23.7	25.2	33.8	35.0	34.1	33.6	34.1	34.4
γ=1	23.1	24.0	21.4	31.9	29.2	30.0	28.9	28.4	29.2

Table 4.1: β_2 powers, 200 patients. Predicted power = 35.2%.

	$(Var(U_1), Var(U_2)) = (1.19, 0.0003)$			$(Var(U_1), Var(U_2)) = (1.19, 0.00003)$			$(Var(U_1), Var(U_2)) = (0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	57.1	56.8	55.1	60.0	60.2	60.1	60.3	60.6	60.1
γ=0.5	48.7	49.6	46.7	59.3	59.6	58.1	59.3	59.0	59.9
γ=0.75	41.6	42.3	45.0	55.6	57.2	58.0	58.5	57.0	58.9
γ=1	33.5	32.1	33.7	51.3	53.0	54.3	51.5	49.3	54.0

Table 4.2: β_2 powers, 400 patients. Predicted power = 60.8%.

	$(Var(U_1), Var(U_2)) = (1.19, 0.0003)$			$(Var(U_1), Var(U_2)) = (1.19, 0.00003)$			$(Var(U_1), Var(U_2)) = (0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	70.1	70.8	73.9	78.1	78.0	76.9	78.0	78.0	77.2
γ=0.5	63.5	67.0	62.3	75.2	75.5	75.2	77.4	75.6	76.3
γ=0.75	55.0	56.7	56.3	72.3	74.4	74.2	74.1	73.6	72.4
γ=1	45.0	45.4	47.3	69.6	68.6	68.9	70.0	69.9	70.4

Table 4.3: β_2 powers, 600 patients. Predicted power = 78.2%.

	$Var(U_1, U_2) = (1.19, 0.0003)$			$Var(U_1, U_2) = (1.19, 0.00003)$			$Var(U_1, U_2) = (0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	83.0	83.1	83.8	88.5	87.6	88.0	88.1	88.4	87.8
γ=0.5	73.0	72.3	75.0	86.8	88.2	88.6	87.1	87.0	88.3
γ=0.75	67.1	67.1	67.3	86.1	84.0	84.4	85.1	87.2	86.5
γ=1	58.0	57.1	56.5	77.7	81.1	79.7	80.7	80.2	82.1

Table 4.4: β_2 powers, 800 patients. Predicted power = 88.5%.

γ Results

Data was simulated and the results of power for different trial scenarios and different γ's were calculated. The results are shown in Tables 4.5 - 4.7. For each table, the calculated power is compared to the estimate from Equation (4.31) using the truncated moments as determined by the data and the estimate from Equation (4.34) using a uniform approximation for the truncated moments. The results show that $Var(U_1)$, $Var(U_2)$ and number of time points all had an effect on the power for γ. In general, greater variances resulted in higher powers for the simulated data. For $Var(U_1, U_2) = (1.19, 0.0003)$, powers were found to be high, with the majority of empirical power being estimated as 100%. Therefore Table 4.7 lists the only power values that were less than 100. In this study, for the simulations where $Var(U_2) = 0.0003$ the empirical powers were found to be substantially higher than when $Var(U_2) = 0.00003$. The power was also greater when $Var(U_1)$ was higher.

The change in power appears to be sensitive to the γ parameter. For the simulated results, in the majority of cases where $\gamma \geq 0.5$ the power from the simulated data was approximately 100%. The exception was when $Var(U_1, U_2) = (0.6, 0.00003)$ with 200 patients in the study, as shown in Table 4.5. Figure 4.4 shows a comparison of the simulated power with the two approximations of power using Equations (4.31) and (4.34) for $\gamma = 0.25$.

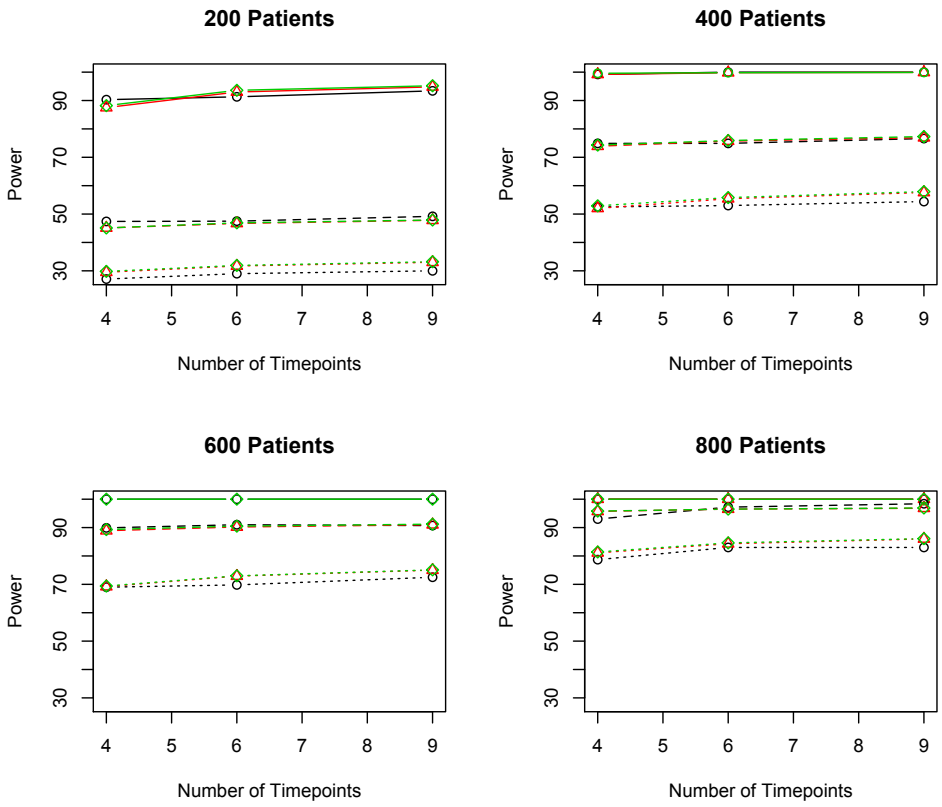


Figure 4.4: Powers for $\gamma = 0.25$. Black = Power from simulations, Red = Power calculated from Equation (4.34) using approximate truncated moments, Green = Power calculated from Equation 4.31 using actual value of truncated moments. The different line types in Figure 4.5 indicate the combinations of $Var(U_1)$ and $Var(U_2)$ where $(1.19, 0.0003)$ is represented by a solid line, $(1.19, 0.00003)$ by dashed lines and $(0.6, 0.00003)$ by dotted lines.

Figure 4.4 demonstrates that as the number of time points increased, there was a slight increase in the power for γ in most cases. The graphs indicate that the approximation formula generally acted as a good predictor for the calculated power in these simulations. In simulations with $Var(U_1, U_2) = (0.6, 0.00003)$, the simulated power was found to be overestimated by the derived formulae in some cases. However, the majority of predicted powers were within 3% of the empirical estimate.

				Power calculated from simulations			Power approximated using estimated moments			Power approximated using calculated moments		
$Var(U_1)$	$Var(U_2)$	γ	patients	$ntms = 4$	$ntms = 6$	$ntms = 9$	$ntms = 4$	$ntms = 6$	$ntms = 9$	$ntms = 4$	$ntms = 6$	$ntms = 9$
0.6	3×10^{-5}	0.25	200	27.1	29.0	30.0	29.5	31.7	33.0	29.8	31.9	33.2
0.6	3×10^{-5}	0.5	200	79.1	81.7	83.0	81.4	84.3	86.0	81.6	84.6	86.2
0.6	3×10^{-5}	0.75	200	98.2	98.2	98.9	98.9	99.0	99.3	99.0	99.2	99.3
0.6	3×10^{-5}	1	200	99.6	99.9	99.8	100	100	100	100	100	100
0.6	3×10^{-5}	0.25	400	52.5	53.0	54.4	52.0	55.4	57.6	52.9	55.8	57.9
0.6	3×10^{-5}	0.5	400	93.0	93.6	96.8	97.9	98.2	98.7	98.1	98.5	98.8
0.6	3×10^{-5}	0.75	400	99.8	99.8	100	100	100	100	100	100	100
0.6	3×10^{-5}	1	400	100	100	100	100	100	100	100	100	100
0.6	3×10^{-5}	0.25	600	69.0	69.8	72.5	69.2	72.9	75.0	69.4	73.0	75.1
0.6	3×10^{-5}	0.5	600	98.4	99.6	99.8	99.8	99.9	99.9	99.9	99.9	99.9
0.6	3×10^{-5}	0.75	600	100	100	100	100	100	100	100	100	100
0.6	3×10^{-5}	1	600	100	100	100	100	100	100	100	100	100
0.6	3×10^{-5}	0.25	800	78.7	83.0	83.0	81.1	84.3	86.0	81.4	84.6	86.1
0.6	3×10^{-5}	0.5	800	99.6	99.8	100	100	100	100	100	100	100
0.6	3×10^{-5}	0.75	800	100	100	100	100	100	100	100	100	100
0.6	3×10^{-5}	1	800	100	100	100	100	100	100	100	100	100

Table 4.5: Power for γ when $(Var(U_1), Var(U_2)) = (0.6, 0.00003)$

				Power calculated from simulations			Power approximated using estimated moments			Power approximated using calculated moments		
$Var(U_1)$	$Var(U_2)$	γ	patients	$ntms = 4$	$ntms = 6$	$ntms = 9$	$ntms = 4$	$ntms = 6$	$ntms = 9$	$ntms = 4$	$ntms = 6$	$ntms = 9$
1.19	3×10^{-5}	0.25	200	47.4	47.5	49.2	45.1	46.7	47.8	45.1	46.9	47.9
1.19	3×10^{-5}	0.5	200	94.8	95.5	96.3	96.7	97	97.3	96.8	97.2	97.4
1.19	3×10^{-5}	0.75	200	100	99.9	100	99.9	100	100	100	100	100
1.19	3×10^{-5}	1	200	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	0.25	400	74.9	74.9	76.6	73.9	75.7	76.9	74.3	75.9	77.3
1.19	3×10^{-5}	0.5	400	99.9	100	100	99.9	99.9	99.9	99.9	99.9	100
1.19	3×10^{-5}	0.75	400	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	1	400	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	0.25	600	89.9	91.0	90.7	88.9	90.2	91.0	89.3	90.5	91.2
1.19	3×10^{-5}	0.5	600	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	0.75	600	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	1	600	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	0.25	800	93.0	97.2	98.4	95.7	96.4	96.8	95.8	96.5	96.9
1.19	3×10^{-5}	0.5	800	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	0.75	800	100	100	100	100	100	100	100	100	100
1.19	3×10^{-5}	1	800	100	100	100	100	100	100	100	100	100

Table 4.6: Power for γ when $(Var(U_1), Var(U_2)) = (1.19, 0.00003)$

				Power calculated from simulations			Power approximated using estimated moments			Power approximated using calculated moments		
$Var(U_1)$	$Var(U_2)$	γ	patients	$ntms = 4$	$ntms = 6$	$ntms = 9$	$ntms = 4$	$ntms = 6$	$ntms = 9$	$ntms = 4$	$ntms = 6$	$ntms = 9$
1.19	3×10^{-4}	0.25	200	90.3	91.3	93.4	87.5	93.0	94.8	88.2	93.6	95.2
1.19	3×10^{-4}	0.25	400	99.2	100	100	99.2	99.8	99.9	99.6	99.8	99.9

Table 4.7: Power for γ when $(Var(U_1), Var(U_2)) = (1.19, 0.0003)$

β_1 Results

The results in Tables 4.8-4.11 show that for almost all γ and variance combinations, higher numbers of longitudinal time points resulted in an increase in the empirical power of the study. This echoes the sample size formula for linear mixed models in Equation (4.35). This was less detectable through the simulations for trials with fewer patients and greater variances. In particular, for the second set of simulations (Table 4.9), there are some cases where a the relationship between time points and power is difficult to detect. However, in Table 4.11, which demonstrate the empirical power for trials with 800 patients total, there is shown to be a strong correlation between these two properties. The value of γ was found to have no effect on the power for β_1. Figure 4.5 shows the mean power across all values of γ for each combination of $Var(U_1)$, $Var(U_2)$ and number of patients plotted against the number of time points used in each simulation.

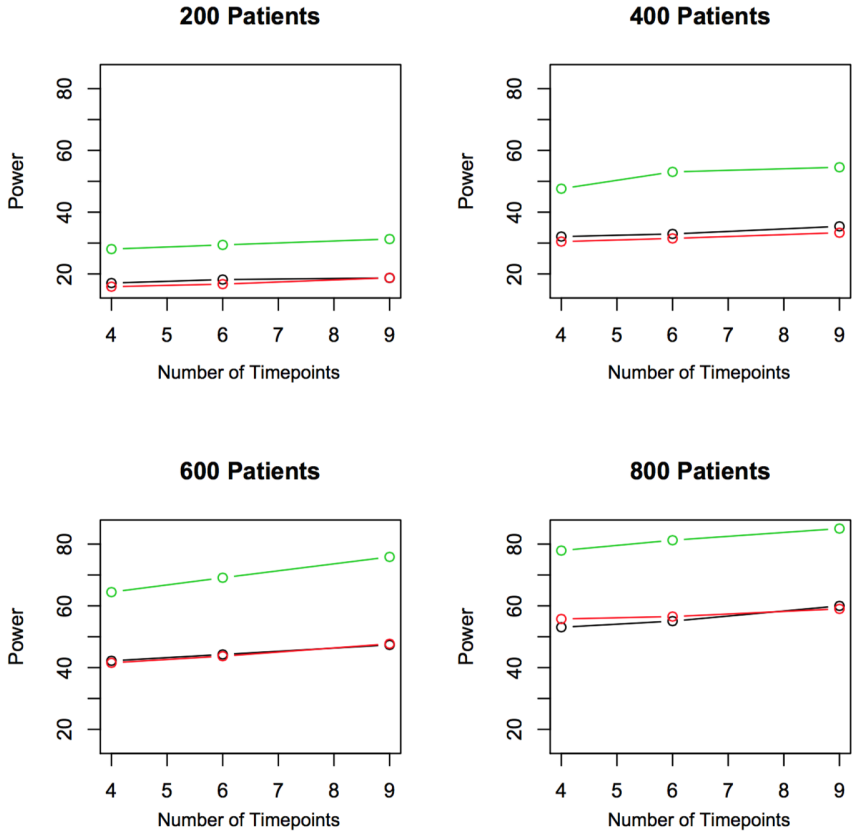


Figure 4.5: Plot of mean β_1 power across $\gamma's$ against $ntms$ by $Var(U_1, U_2)$ combination. $Var(U_1, U_2) = (1.19, 0.0003)$(black), $Var(U_1, U_2) = (1.19, 0.00003)$(red), $Var(U_1, U_2) = (0.6, 0.00003)$(green)

Figure 4.5 indicates that when $Var(U_2)$ is increased, this has little or no effect on the power for β_1. However, this is not the case for $Var(U_1)$, which results in a decrease in power for higher variances. This decrease can be observed for all 4 sets of simulations with different numbers of patients. The final row of Tables 4.8 - 4.11 show the results of the powers calculated using only a complete case analysis. The powers when using joint modelling were found to be higher than for a complete case analysis in all simulations. Figure 4.6 shows the mean increase in power when compared to a complete case analysis plotted against number of timepoints for all simulation properties.

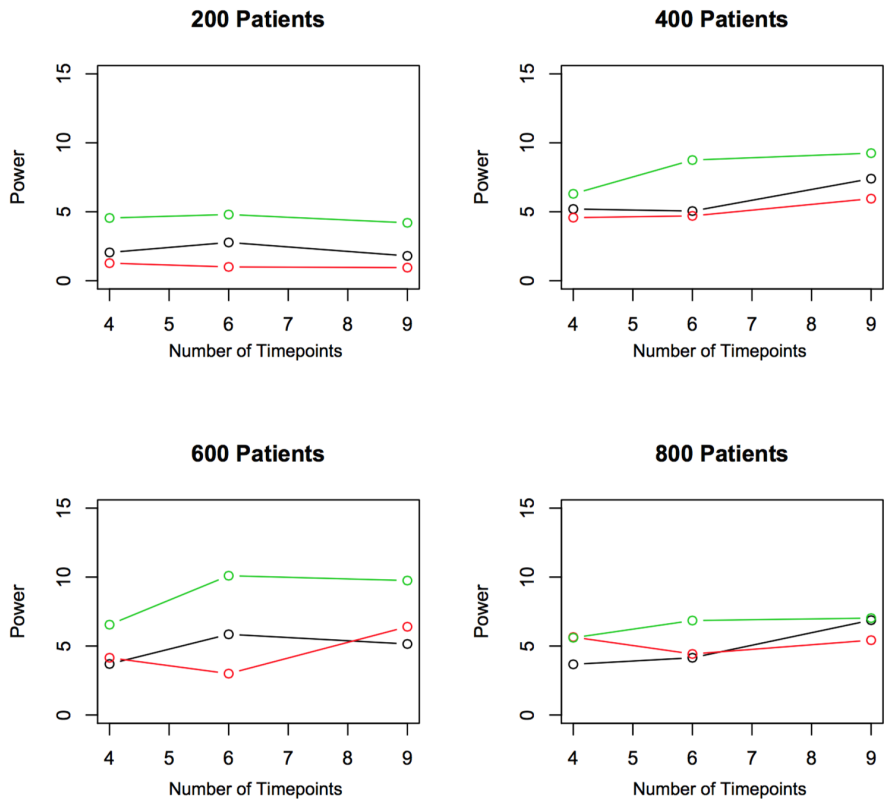


Figure 4.6: Increase in power for β_1 compared to the powers for a complete case analysis. $Var(U_1, U_2) = (1.19, 0.0003)$(black), $Var(U_1, U_2) = (1.19, 0.00003)$(red), $Var(U_1, U_2) = (0.6, 0.00003)$(green)

In general, there was a greater increase in power for larger values of $Var(U_1)$. On average there appeared to be a greater increase in power for larger number of time points, although this was not the case for all simulations. In the majority of simulations, the power when using joint modelling was at least 4% higher than for a complete case analysis.

$(Var(U_1), Var(U_2)) =$	$(1.19, 0.0003)$			$(1.19, 0.00003)$			$(0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	17.8	16.4	17.5	15.6	14.6	18.6	26.4	27.6	28.8
γ=0.5	16.7	20.4	20.5	14.8	15.7	19.2	28.7	31	32.4
γ=0.75	16.3	18.2	18.2	16.3	16.4	17	28.8	28.6	32.3
γ=1	17.4	17.7	18.6	16.8	20.1	20.2	28.3	30.4	31.7
Complete Case	15.0	15.4	16.9	14.6	15.7	17.8	23.5	24.6	27.1

Table 4.8: Empirical Powers for 200 Patients

$(Var(U_1), Var(U_2)) =$	$(1.19, 0.0003)$			$(1.19, 0.00003)$			$(0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	30.8	34.7	35.5	31.4	30.4	32.2	46.8	53.5	55.4
γ=0.5	31.2	34.7	36.1	29.2	34	34.7	46	52.1	54.8
γ=0.75	33.6	31	34.8	30.3	30.8	32.1	48.4	52.1	52.8
γ=1	32.8	31.4	35.2	31	30.8	34.4	49.2	54.5	55.2
Complete Case	26.9	27.9	28.0	25.9	26.8	27.4	41.3	44.3	45.3

Table 4.9: Empirical Powers for 400 Patients

$(Var(U_1), Var(U_2)) =$	$(1.19, 0.0003)$			$(1.19, 0.00003)$			$(0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	44	44.6	48.8	41.0	45.6	45.8	62.2	69.2	76
γ=0.5	41.5	44.9	47.9	40.6	45.0	48.6	61.4	69.6	74.4
γ=0.75	41.3	44.1	46.7	39.8	42.5	48	66.2	68	74.4
γ=1	42.0	43.4	46	44.8	41.7	48.4	68.0	69.6	78.6
Complete Case	38.5	38.4	42.2	37.4	40.7	41.3	57.9	59.0	66.1

Table 4.10: Empirical Powers for 600 Patients

$(Var(U_1), Var(U_2)) =$	$(1.19, 0.0003)$			$(1.19, 0.00003)$			$(0.6, 0.00003)$		
	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9	ntms=4	ntms=6	ntms=9
γ=0.25	52.1	53.1	59.0	54.8	55.4	57.8	77.9	81.8	85
γ=0.5	53.2	56.1	58.6	56.2	57.2	59.1	78.6	79.8	85.8
γ=0.75	52.9	54.2	61.3	54.4	57.3	61.2	79.0	81.2	84.1
γ=1	54.1	56.8	61	57.6	56.2	58	76.1	82.2	85.2
Complete Case	49.4	50.9	53.1	50.1	52.1	53.6	72.3	74.4	78.0

Table 4.11: Empirical Powers for 800 Patients

Simulated Power For MAGNETIC

We estimate the power within the MAGNETIC trial for each parameter. The primary outcome for MAGNETIC was the difference in treatment after 60 minutes, for which the power was specified as 80% in the protocol. Data was simulated with 508 patients, 27% dropout, $Var(U_1) = 1.19$ and $Var(U_2) = 0.00003$ to go with the MAGNETIC trial analysis. From the Henderson et al. specification of the joint model, longitudinal data was generated for 7 different time points, namely $t = 0, 20, 40, 60, 120, 180, 240$.

Using bootstrapping methods for the 95% confidence interval and 1000 simulations, the power for β_1, β_2 and γ were calculated. The power for β_1 based on the simulated data was found to be 43.4%. The corresponding powers for β_2 and γ were 72.5% and 74.6% respectively. The predicted powers based on formulae (4.37) and (4.34) were 72.0% for β_2 and 77.2% for γ, which indicated a reasonably successful approximation.

4.6 Discussion

In this chapter, the powers for β_1, β_2 and γ parameters were investigated. A sample size formula for γ and approximate sample size formula for β_2 was derived for the random slope and intercept joint model. The sample size for γ was found to be dependent on a number of factors, including the pre-specified power, the number of time points and the covariance matrix of U. Due to the nature of the model, knowledge of the variances of these random effects are required to derive the sample size formula in an RCT. This may seem problematic, however, in many trials preliminary studies are carried out and these can be used to approximate $Var(U_1)$ and $Var(U_2)$ prior to a larger scale trial. The simulation studies in this chapter showed that these formulae provide accurate estimation of power. Also, the proposed approximations for the truncated moments appeared to perform well for estimating power.

In the case of β_2 for time-to-dropout, when $\gamma = 0$, the power is only related to the magnitude of β_2, the proportion of patients in each treatment group and the number of dropouts. As γ increases for random slope and intercept models, the power decreases. Also when $\gamma \neq 0$, the power decreases for higher values of $Var(U_2)$, although the power is unrelated to $Var(U_1)$. To account for the loss in power, to calculate the power for β_2 we can combine Equation (4.15) with a variance inflation factor, as shown in Section 4.5.2. The relationship has been verified for the MAGNETIC trial scenario with a follow up time of 240, however more work has to be done to investigate the relationship for different trial designs.

When the parameter of interest is the longitudinal treatment effect, β_1, using joint models over the complete case analysis will increase the power in a trial and therefore require fewer patients. Likewise increasing the number of time points will have a positive effect on power. Despite this clarification, with all the different factors to consider, the best method of calculating the power for β_1 is to use simulation methods.

In practice, the formulae proposed in this chapter will provide a useful aid for trial design when using joint modelling. However, in order to accurately estimate the power for β_2 and γ, a more concise definition of γ must be available and the properties of this parameter should be understood. In the next chapter, a visualisation of γ is provided and a detailed examination of the relationship between γ and longitudinal outcome pre-dropout carried out.

Chapter 5

A Visualisation of the γ Parameter

5.1 Introduction

The majority of the work in this thesis has been based around a random slope and intercept joint model, which utilises a latent Gaussian process to model the progression of a longitudinal outcome over time alongside dropout. In Chapter 4, a sample size formula was generated for the γ association parameter. However, currently knowledge about the properties of γ is limited.

It has been established that RCT analyses with a negative γ estimate contain patients that are more likely to dropout with a lower longitudinal reading, while a positive γ estimate indicates that patients with a higher longitudinal reading are more likely to dropout [10]. Although this information is useful, it makes no comment about how the magnitude of γ is related to the change in longitudinal outcome of patients prior to dropout. Currently no details of how to interpret the magnitude of the γ parameter has been described in clinical literature, however knowledge of the relationship between the outcome of patients pre-dropout and γ can be informative in a clinical environment.

Using two simulation studies, the aim of this chapter is to provide a visualisation of the relationship between the magnitude of change in longitudinal outcome pre-dropout and different values of γ. The first study investigates the mean dropout profiles for various γ's based upon the parameters of the MAGNETIC trial. The second study aims to quantify the relationship by using simulated data based upon a wider range of trial properties and simulation parameters. A simulation study that focuses on the relationship between γ and the longitudinal profiles of patients prior to dropout has not been previously carried out in statistical literature.

Information provided in Chapter 4 displays a need for greater understanding and context to be given to the γ parameter in joint models. This study aims to provide a useful

overview of the properties of γ for statisticians and clinicians employing a random slope and intercept joint modelling analysis.

5.2 Mathematical properties of the γ parameter

A definition of the random slope and intercept joint model is provided in Section 1.4.3, with the longitudinal and time-to-event elements of the model defined in Equations (1.1) and (1.2) respectively.

In this model, there are no multiplicative interactions considered between the fixed and random effects. Differentiating the longitudinal element of the model yields:

$$\frac{dy_i}{dt} = \alpha + U_{2i}. \tag{5.1}$$

Therefore the change in longitudinal outcome for each patient is only estimated by α and the subject specific slope random effect. For this simulation study, the focus will be models with no fixed time variable (α). In this instance, change in longitudinal outcome over time is only dependent on the U_2 random slope parameter. In this case, U_{2i} can be defined as the mean increase in longitudinal outcome for each increase in unit time for patient i.

Therefore, for the aforementioned specification, we can substitute $U_{2i} = y_i'$ into the time-to-event component of the model:

$$\lambda_i(t) = \lambda_0(t) \exp\{x_{2i}(t)'\beta_2 + \gamma(U_{1i} + y_i't)\} \tag{5.2}$$

In the semi-parametric Cox-Model, the baseline hazards are uniquely calculated, and are different depending on the trial dropout characteristics [112]. If the baseline hazard

function is constant, the formula above indicates that as time increases by one unit, the hazard for that patient increases by a proportion of $e^{\gamma y_i'}$. This means that the higher the value of γ, the more likely a patient with higher outcomes (values of $U_1 + U_2 t$) are to dropout. Similarly, a negative γ indicates patients with decreasing outcomes are more likely to dropout. As the nature of dropout is changing on an exponential scale, γ is sensitive. For example a γ value of 5 will result in patients with the same positive values of U_1 and U_2 being $\frac{e^{5(U_1+U_2)}}{e^{U_1+U_2}} = e^{4(U_1+U_2)}$ times more likely to dropout as time progresses to t=1, than if $\gamma = 1$ and all other properties were the same. As another illustrative example, if U_1 and U_2 were both equal to 1, this would equate to a patient being approximately $e^5 \approx 148$ times more likely to drop out at $t = 1$ than at $t = 0$.

As the both longitudinal and time-to-event outcomes are dependent on some random effects U_1 and U_2, the properties of the distributions of the latent variables have a direct impact on the estimated value of γ.

Due to the finite number of patients in a study and discrete time points, for higher variances there are inevitably values of γ which fail to make clinical or statistical sense. For example it may be difficult to estimate in a trial of 500 people that a patient is $e^{10} \approx 22,026$ times more likely to leave a study as time increases by a unit time interval (i.e. when $\gamma = 10$), as there may not be enough patients to accurately determine this value. Calculating the exact continuous mean profiles for patients that dropped out would require an infinite number of time points, which is understandably not practical. Therefore, the more time-points and patients available, the more accurate the estimate of γ. Also, the Gaussian and mean 0 nature of U_1 and U_2 will restrict the possible longitudinal values pre-dropout for combinations of γ, $Var(U_1)$ and $Var(U_2)$ as in the formula proposed by Henderson et al [10], there is a multiplicative relationship between γ and a combination the random Gaussian variables.

In order to understand the magnitude of γ, the profiles of patients that dropped out, the variance of U_1, U_2 and the percentage of dropout must all be taken into consideration. The last of these is largely defined by the baseline hazard.

5.3 Simulation Study 1 - Methods and Results

The aim of simulation study 1 is to obtain a visual appreciation for the relationship between change in outcome pre-dropout and γ parameter. For the first simulation study, data is generated for a longitudinal outcome that also records an event. The event-time in this simulation study is defined as dropout to remain consistent with the rest of this thesis. Data is generated using the random slope and intercept Henderson joint model specification outlined in Chapter 1.4.3, using similar techniques as in Chapter 2.4.1. Dropout was generated from the Cox-proportional hazard based event-time and then transformed to the last time point which was non-missing. As in the previous simulation study, the baseline hazard was generally set to be flat unless stated otherwise and continuous longitudinal outcome data was used.

In simulation study 1 datasets are generated by varying the γ parameter while β_1, β_2, $Var(U_1)$, $Var(U_2)$ and the percentage of dropout are fixed. For each combination of trial properties, 1000 data sets are simulated. Data is generated based on the two arm MAGNETIC trial in order to discover and highlight the differences in longitudinal profiles prior to dropout depending on γ.

For this simulation, $\beta_0 = 5.613$ as the intercept and $\beta_1 = -0.2$ are fixed, which were the approximate parameter estimates for the MAGNETIC data. Similarly the variances of U_1 and U_2 were fixed at 1.19 and 0.00003 respectively. We set $\beta_2 = 0.5$, dropout equal to 20% by varying the flat baseline hazard and simulations are generated for $\gamma = 0, 0.25, 0.5, 0.75$

and 1. There was no necessity to simulate data for negative γ as this would result in the same magnitude of change in outcome prior to dropout as the corresponding positive value, but with decreasing profiles. The error variance was set to 0.5.

The review in Chapter 3 demonstrated that 92% of the sample of the balanced longitudinal RCTs had less than 8 time points recorded [60]. Therefore, simulated datasets are generated for the common trial scenarios of 4,5 and 6 time points as well as 9 longitudinal time points. The MAGNETIC trial had a follow up time of 240. To mirror this, time points are generated by dissecting this into $ntms = 4, 5, 6, 9$ equal intervals. 1000 data sets are simulated for each combination of γ and number of time points for 500 patients per group. Comparisons are made by observing the backwards longitudinal dropout plots for all patients that left the study, with a view to identifying the key areas of variation in the dropout profiles. This way the general trends of patients just before they left the study could be clearly visualised.

Figure 5.1 shows the mean longitudinal profiles prior to dropout for the simulations where $\gamma = 0, 0.25, 0.5, 0.75, 1$ for 4,5,6 and 9 timepoints respectively.

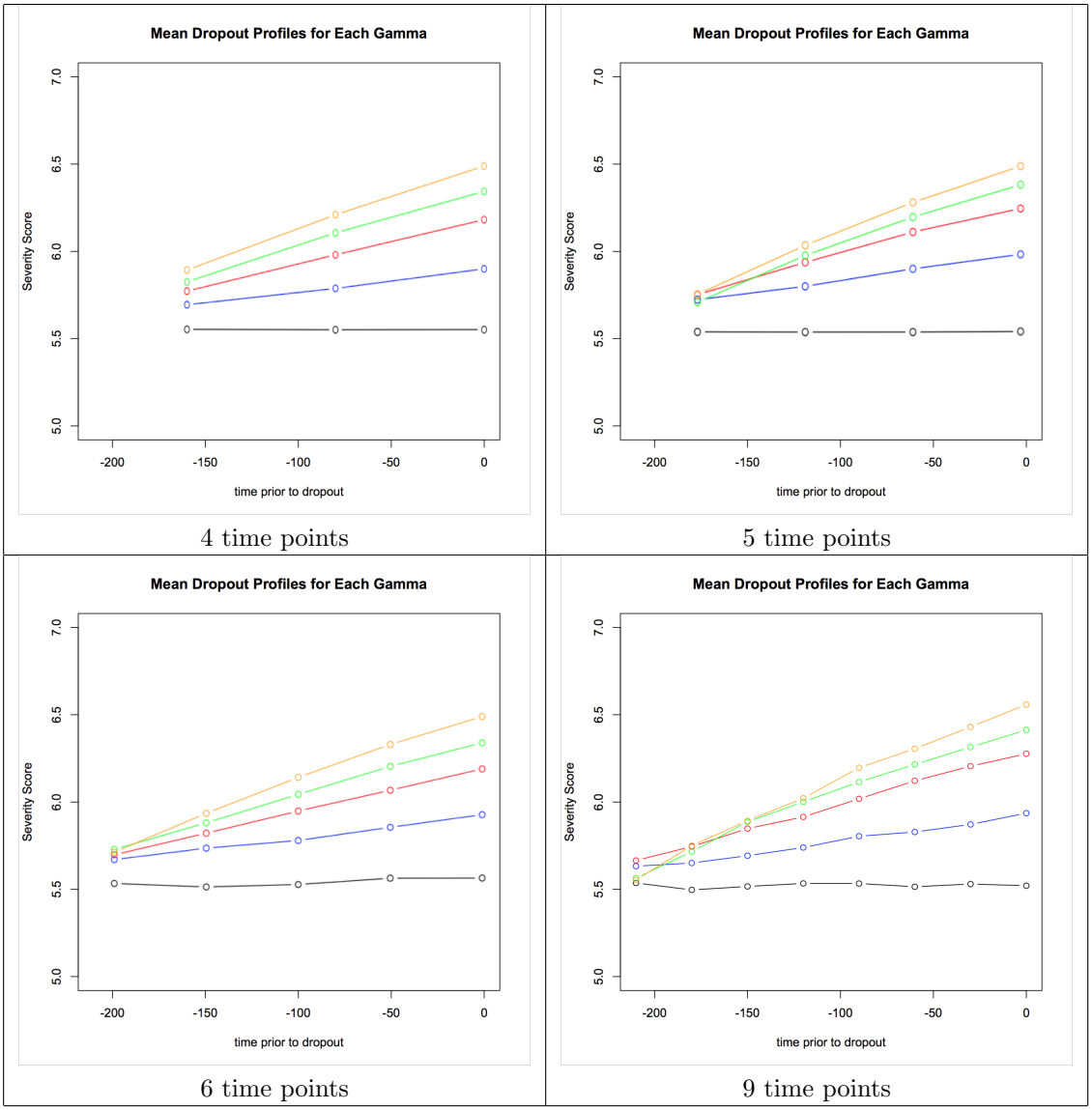


Figure 5.1: Longitudinal dropout profiles based on simulations; $\gamma = 0$ (black), $\gamma = 0.25$ (blue), $\gamma = 0.5$ (red), $\gamma = 0.75$ (green), $\gamma = 1$ (yellow), Dropout = 20%

Figure 5.1 indicates that the magnitude of increase in longitudinal outcome prior to dropout is proportional to the value of γ. When $\gamma = 0$, there is no discernible change in patient profiles before leaving a study, while all other mean profiles show an increase over time. The graphs representing simulations with a higher number of timepoints had greater levels of longitudinal profile variation for different γ values in the moments leading up to dropout. For example, in the plot for $ntms = 9$ there is very little difference between the

dropout profiles shown for $\gamma = 0.75$ and $\gamma = 1$ before $t = -90$. The difference between the dropout profiles is also more evident for lower values of γ, with the difference in slope between $\gamma = 0$ and $\gamma = 0.25$ being larger than between $\gamma = 0.75$ and $\gamma = 1$.

One key area of variation that should be investigated further is the difference in longitudinal outcome in the last 2 timepoints leading up to dropout, for different *ntms*. In Figure 5.2 the differences between the last two time points for each group of simulations are plotted.

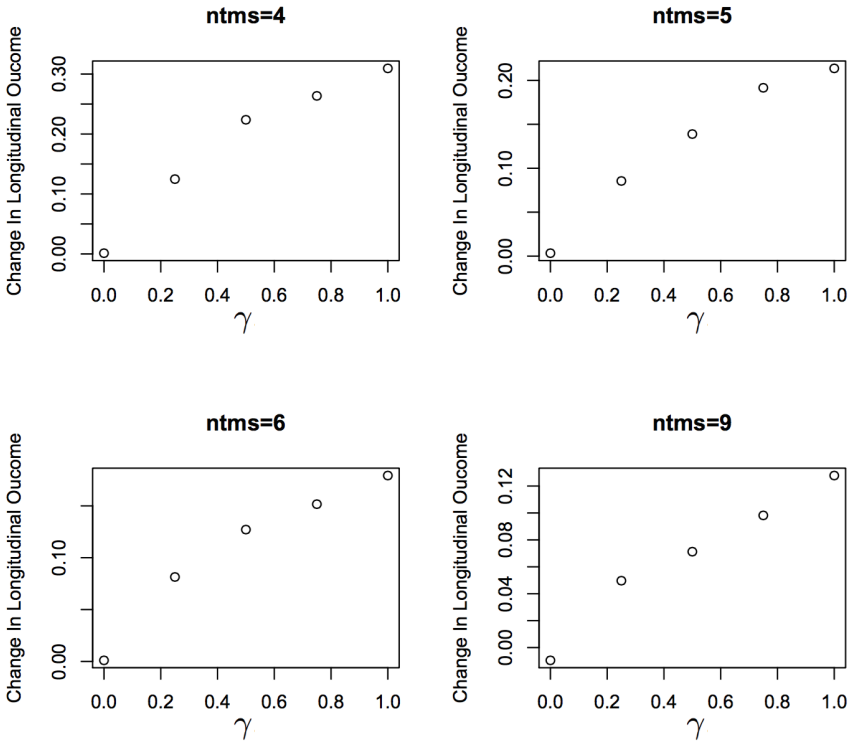


Figure 5.2: Difference in mean profile between last two time points

Figure 5.2 appears to show a clear pattern when observing longitudinal outcome change immediately prior to dropout for different γ values. The graphs demonstrate a larger difference in the change in longitudinal outcome between $\gamma = 0$ and $\gamma = 0.25$, than for higher values of γ. As γ increased the differences in change in longitudinal outcome

decreased. This was the case for all variations of different timepoints. To compare the results from each simulation for varying *ntms*, the normalised mean change in longitudinal outcome per unit time was calculated by dividing the change in longitudinal outcome by the distance between these last two points.

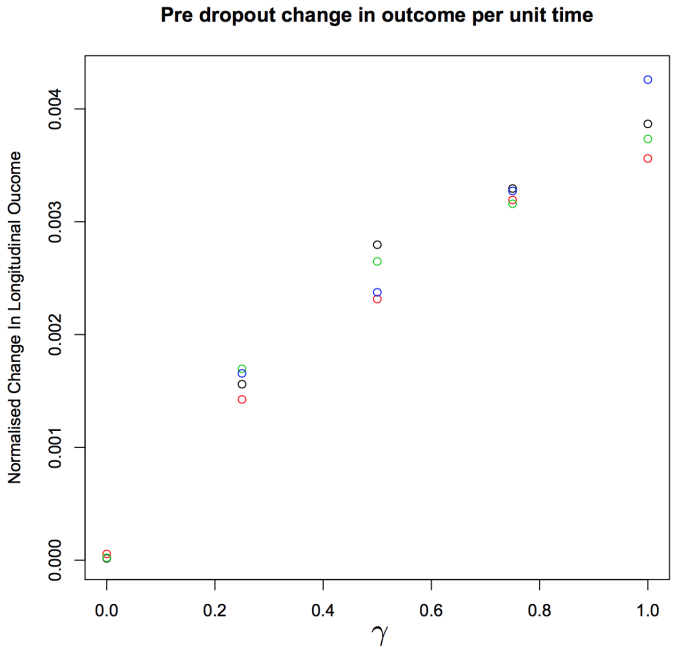


Figure 5.3: Difference in mean profile between last two time points prior to dropout. The different coloured dots represent the simulated data for different numbers of time points; ntms = 4 (black), ntms = 5 (red), ntms = 6 (green), ntms = 9 (blue)

The graph in Figure 5.3 represents the mean change in longitudinal outcome per unit time for each γ for the last two recorded timepoints. The mean profile immediately prior to dropout is approximately the same for the majority of simulations with different *ntms*. However, a higher mean change in outcome is observed in the simulation with $ntms = 9$ and $\gamma = 1$. While this difference has a magnitude of approximately 0.0002, it is important to note that accurately estimating higher values of γ may require greater amounts of information, and higher numbers of time points will result in more accurate γ estimates. This difference also suggests that the normalised change in outcome is dependent on *ntms*

To provide a greater context to the simulations carried out in this section, confidence intervals of the longitudinal dropout profiles were calculated with patients $n = 100, 200, 400, 500$ per treatment group. 1000 simulations were generated with $ntms = 5$ for each number of patients and the 95% C.I.s calculated. Figure 5.4 shows the confidence intervals for the mean dropout profiles for varying sample sizes and 5 time points.

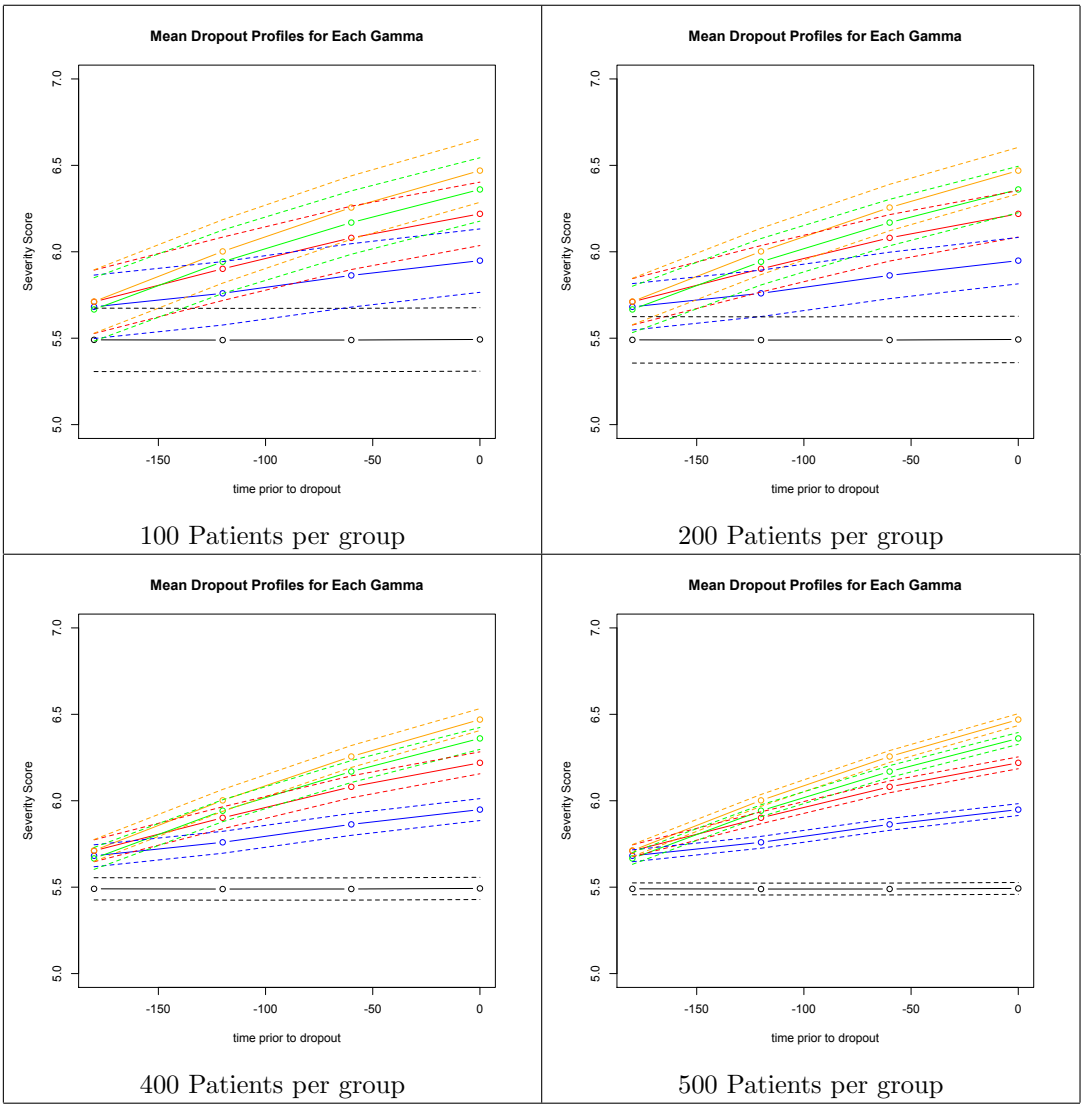


Figure 5.4: Dropout profiles with confidence intervals for different values of γ and numbers of patients per treatment group; $\gamma = 0$ (black), $\gamma = 0.25$ (blue) , $\gamma = 0.5$ (red), $\gamma = 0.75$ (green), $\gamma = 1$ (yellow)

For the simulations with 400 and 500 patients in each group (160 and 200 dropouts), there is a significant difference between the dropout profiles for each value of γ. In simulations with 100 and 200 patients per group (40 and 80 dropouts) there was some overlap in the mean confidence intervals, although the overall mean profiles were very similar for each number of patients.

In this study, visualisations of different γ parameters in terms of change in longitudinal outcome are presented based upon the framework of MAGNETIC. In summary, these simulations identified that the greater the change in mean longitudinal outcome prior to dropout, the higher the estimate of γ. Also, there is a greater difference in longitudinal dropout profiles between smaller values of γ. This difference can be observed by looking at change in longitudinal outcome between the last two time points prior to dropout, however more accurate estimates of γ are given in RCTs with a higher number of timepoints. Traditionally for this type of statistical modelling, the majority of parameters included in the model to describe the outcome variable have no interaction with time. Therefore, focusing on the change in mean longitudinal outcome over time for dropouts may be the most successful approach for approximating γ.

5.4 Simulation Study 2 - Methods and Results

In the second simulation study, the aim is to gain a greater understanding of this relationship between y' and γ by allowing more flexible simulation settings. To explore the properties of different γ values, data is simulated for the simplistic scenario of a trial with longitudinal data measured at 5 different time points at unit intervals, with a view to quantifying the relationship between γ and change in longitudinal outcome pre-dropout for this trial design. This study should provide a basis for future investigations into the properties of γ for alternative trial scenarios.

Simulation study 1 provides a simple visualisation for dropout profiles, however this only investigated the profiles for fixed variances of U_1, U_2 and a fixed percentage dropout. In the second simulation study, the variances of U_1, U_2 and the percentage of dropout will also be varied along with the γ parameter. The initial focus of the study will be on joint modelling data with a flat baseline hazard.

Before considering this new trial scenario, an exploratory analysis is carried out to observe the relationship between $Var(U_2)$ and the dropout profiles for fixed γ, using the MAGNETIC parameters from simulation study 1. Joint longitudinal and dropout data is simulated with $\gamma = 0.5$, dropout percentage equal to 20%. $Var(U_2)$ is varied and the dropout profiles are plotted. For all simulations in this study, the error variance was set to 0.5. The corresponding dropout profiles are shown in Figure 5.5;

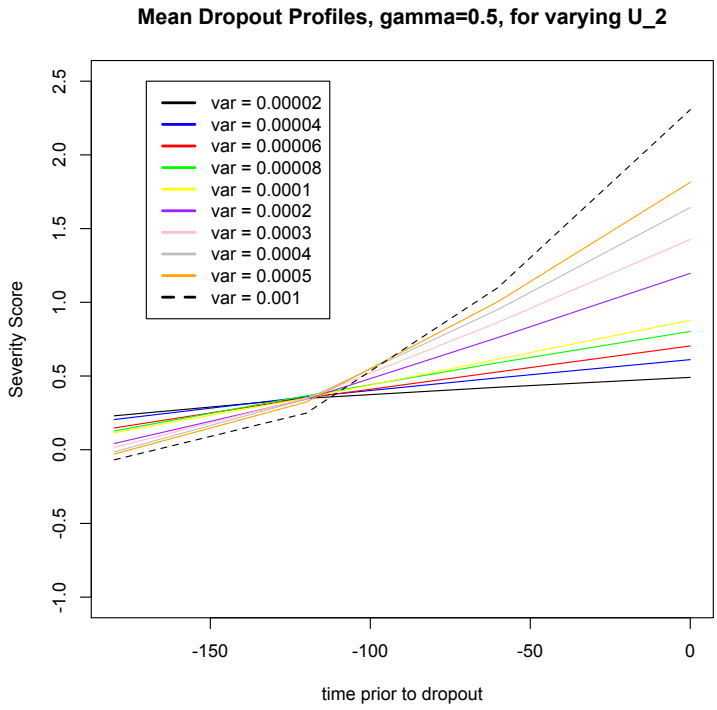


Figure 5.5: Mean dropout profiles for different values of $Var(U_2)$ when $Var(U_1) = 1.19$, dropout percentage = 20% and $\gamma = 0.5$

Figure 5.5 shows that there is a relationship between magnitude of y' for drop outs and $Var(U_2)$ when γ is fixed. For larger values of $Var(U_2)$, there is a greater change in longitudinal outcome pre-dropout. This implies that for the same y', the parameter estimate for γ becomes smaller as $Var(U_2)$ increases. By the model definition, if γ was negative these results would be mirrored, as joint modelling concerns a multiplicative relationship between γ and U_2.

We now attempt to model the relationship between change in outcome pre-dropout and γ for the trial scenario where there are 5 different time points at unit intervals. To quantify change in outcome pre-dropout, we define this as the difference in mean dropout profile between the last two time points leading up to dropout. Data is again simulated from the linear mixed model and Cox-proportional hazards model for the longitudinal and dropout

elements respectively. To assess the raw properties of the γ parameter, it is important to ensure this simulation study's results are as general as possible. By setting $\beta_0 = 0$, $\beta_1 = 0$ and $\beta_2 = 0$ in the simulations, a more general solution can be obtained for the relationship between γ and y'. For all simulations, 1000 data sets are generated with 500 patients in each treatment group. After plots have been generated to investigate the relationship, by using linear modelling techniques to generate approximate mathematical formulae we can establish the extent to which $Var(U_1)$, $Var(U_2)$ and change in outcome affect γ. This will provide a basis for understanding this relationship when considering other trial scenarios. Once a formula has been established, standard diagnostic techniques are used to check the validity of the model.

The effect of varying U_1 on change in longitudinal outcome, y', is now investigated using our newly described trial scenario. We set $Var(U_2) = 0.25$, $\gamma = 0.5, 1$, percentage dropout = 20% and simulate for different values of $Var(U_1)$. Two plots are shown in Figure 5.6; the first is the difference in the mean longitudinal outcomes at 1 time-point before dropout and at dropout, while the second is a plot of the mean change in outcome between the last 2 longitudinal readings pre-dropout when excluding patients that dropped out at $t = 0$. This was to test whether high baseline values for patients that dropped out immediately were distorting our interpretation of y'.

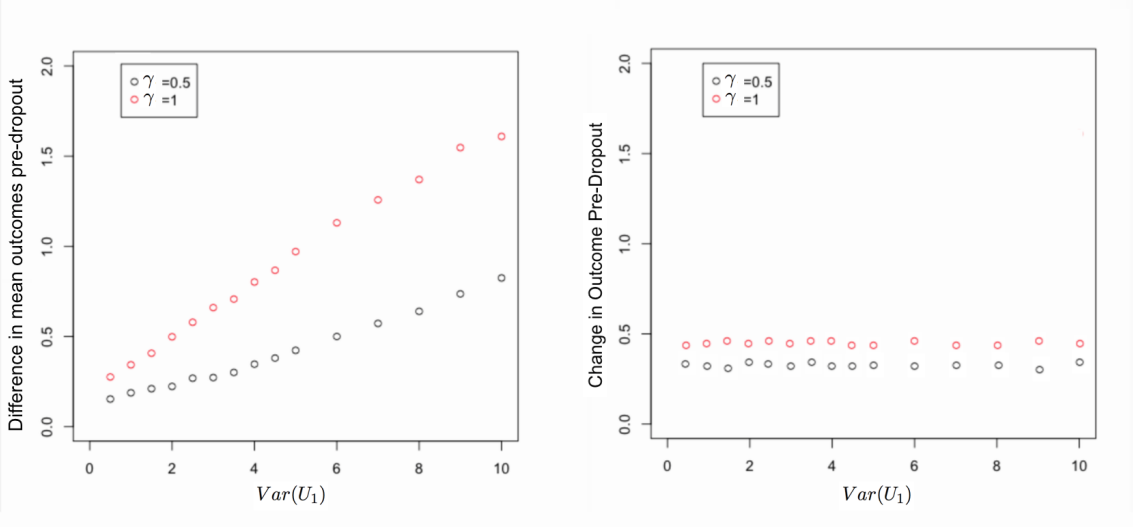


Figure 5.6: Difference in longitudinal profiles pre dropout for different $Var(U_1)$'s including patients that dropped out at t=0 (left) and disregarding those patients (right)

The left panel of Figure 5.6 represents the difference in the last two longitudinal outcomes pre dropout from all patients including those that dropped out at t=0. This graph appears to show an increase in change of longitudinal outcome for higher variances of U_1. However these results are clearly skewed by the patients that dropped out after only one longitudinal reading, as the graph on the right indicates no significant difference in change of outcome pre-dropout when these patients are excluded from the analysis.

To ensure that the influence of $Var(U_1)$ does not provide misleading results, henceforth all graphs and references to graphs of mean dropout profiles exclude the patients that dropped out at $t = 0$. We denote the mean difference in longitudinal value between the last 2 timepoints pre-dropout as δy. To provide a visualisation of the mean dropout profiles with $t = 0$ dropouts excluded, Figure 5.7 shows the simulated mean dropout profiles for varying γ's with $Var(U_1) = 1$, $Var(U_2) = 1$ and percentage dropout equal to 20%.

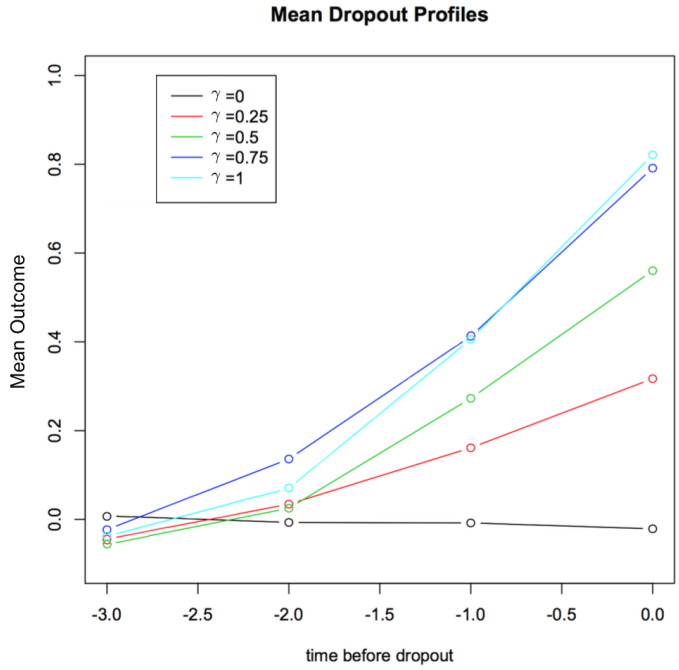


Figure 5.7: Mean dropout profiles with with $Var(U_1) = 1$, $Var(U_2) = 1$ and percentage dropout equal to 20%.

After observing the relationship between $Var(U_2)$, γ and y', the next stage is to quantify this relationship. The percentage of dropout is set as $10\%, 20\%, 30\%, 40\%$ and 50% and 1000 simulated dataset are generated for each combination of γ, $Var(U_2)$ and dropout percentage. $Var(U_1)$ is set equal to 1 for all simulations as this will have no impact on δy. The difference in mean longitudinal outcome between the last two time points pre-dropout is calculated for each outcome. Figure 5.8 shows plot of δy against $Var(U_2)$ for each γ.

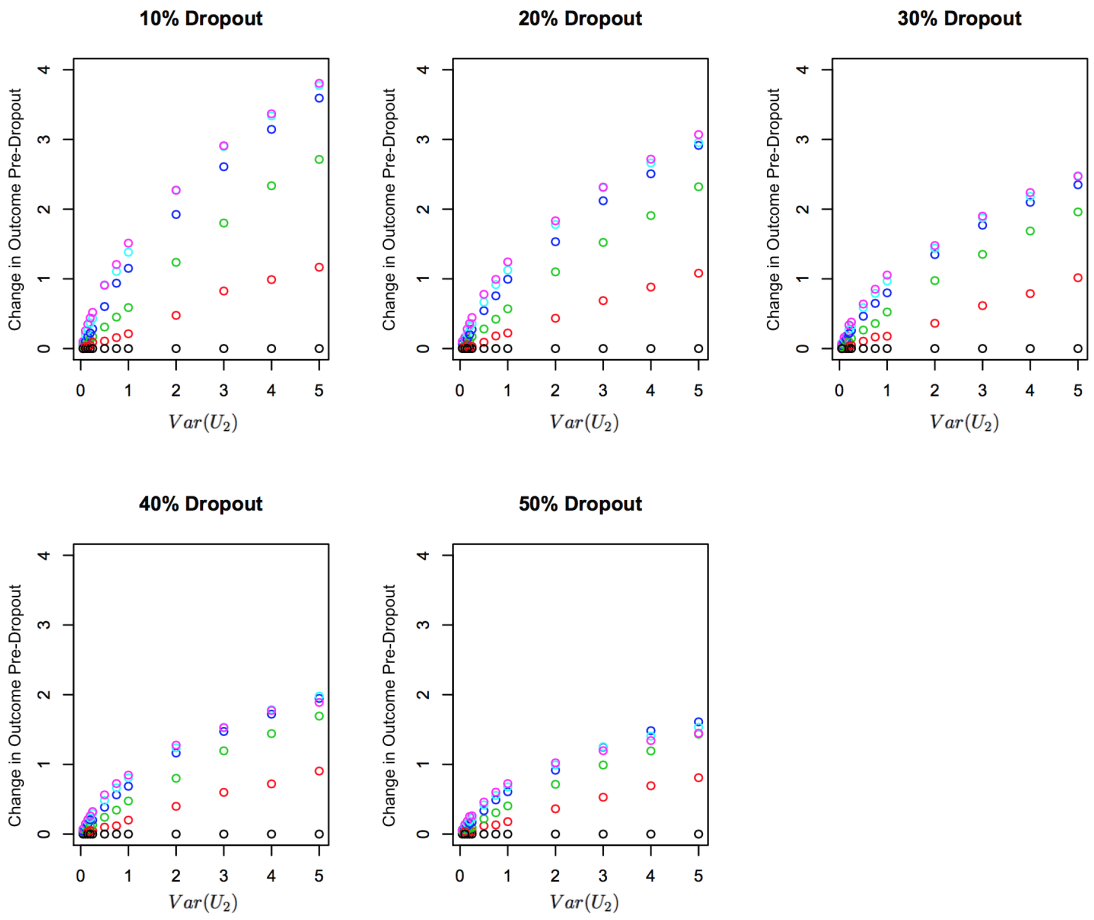


Figure 5.8: Plots of δy against $Var(U_2)$. gamma=0 (black), gamma=0.1 (red), gamma=0.25 (green), gamma=0.5 (dark blue), gamma=0.75 (light blue), gamma=1 (purple)

For simulations with $\gamma = 0$, the change in longitudinal outcome is approximately 0 for all values of $Var(U_2)$. This follows from the formula in which longitudinal outcome and dropout time are unrelated if $\gamma = 0$. When $\gamma \neq 0$, the change in longitudinal outcome increases as the variance of U_2 increases. However, the magnitude of this increase becomes smaller for higher variances. The higher the value of γ the greater the reduction in outcome change as $Var(U_2)$ increases. It would also appear that this deceleration is more noticeable for higher percentages of dropout.

For the simulations with 40% and 50% dropout, the values of δy were similar for higher γ, $Var(U_2)$. As the percentage of dropout was increased, δy decreased for fixed γ and $Var(U_2)$. This demonstrates that the relationship between δy and γ is dependent on the baseline hazard.

Due to γ and $Var(U_2)$ working on an exponential scale in the time-to-event component of the joint model, it can be difficult to identify a difference in longitudinal outcome pre-dropout as there are only a finite number of dropout points and a fixed level of dropout. For a reasonable amount of patients and time points recorded in a study, there will always be an upper threshold which γ can not exceed without the inclusion of unrealistic trial properties. This is particularly evident in the simulations with higher values of γ.

Figure 5.9 presents the simulation results in terms of change in longitudinal outcome plotted against γ.

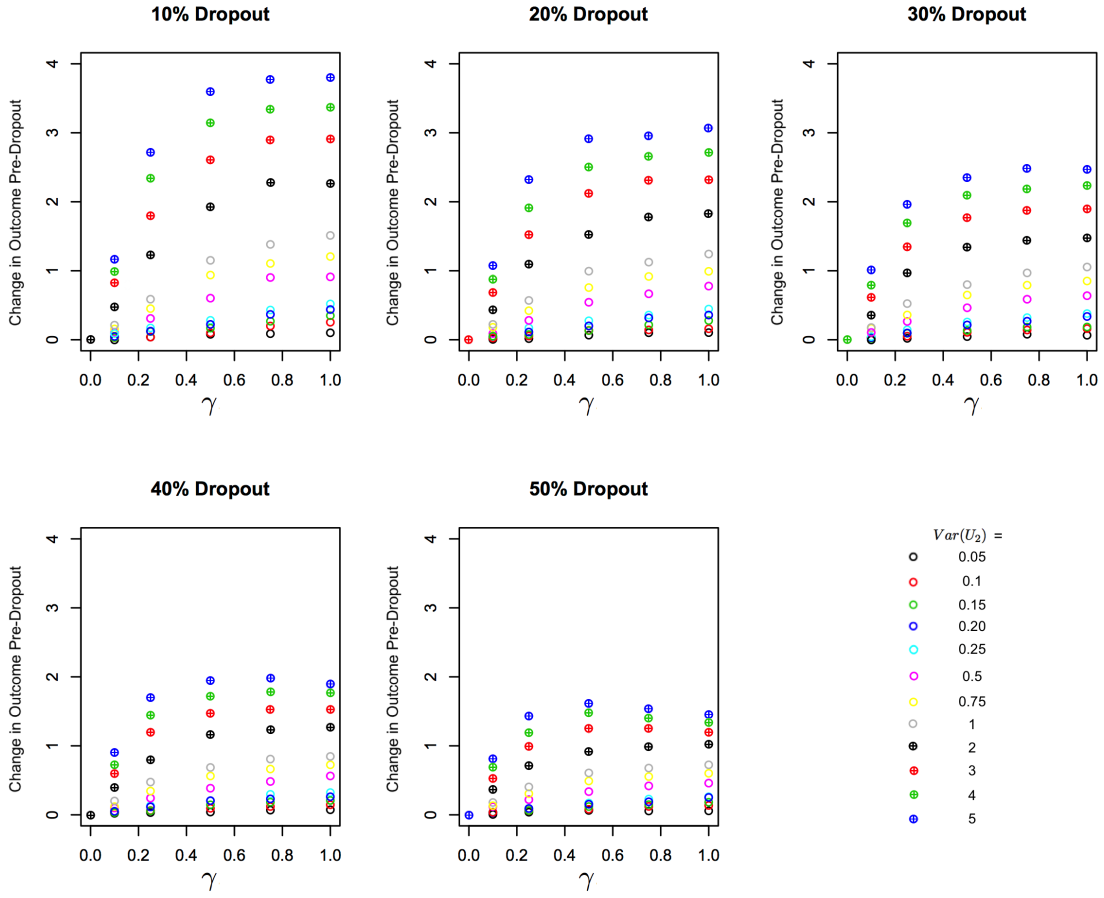


Figure 5.9: Plots of δy against γ.

In general, when $Var(U_2)$ and dropout percentage was fixed, there was an increase in longitudinal outcome as γ increased. However, by pinpointing the results where this is not the case, simulations that used inappropriate combinations of γ, $Var(U_2)$ and percentage dropout can be identified. This was less problematic in the simulations with 10 and 20% dropout, however the following combinations provided misleading results due to the aforementioned factors:

% Dropout	γ	$Var(U_2)$
30%	1	5
40%	1	4
40%	1	5
50%	1	3
50%	0.75	4
50%	1	4
50%	0.75	5
50%	1	5

Table 5.1: Situations where change in longitudinal outcome is difficult to estimate accurately

When aiming to find a relationship between γ, change in outcome and $Var(U_2)$ for this trial design, it is important to find a model which identifies the results of these particular simulations as underestimated.

Figures 5.8 and 5.9 show that there is an interaction between γ and $Var(U_2)$ when modelling δy. To investigate this further, Figure 5.10 shows plot of change in δy against $\gamma \times Var(U_2)$ with different γ values identified.

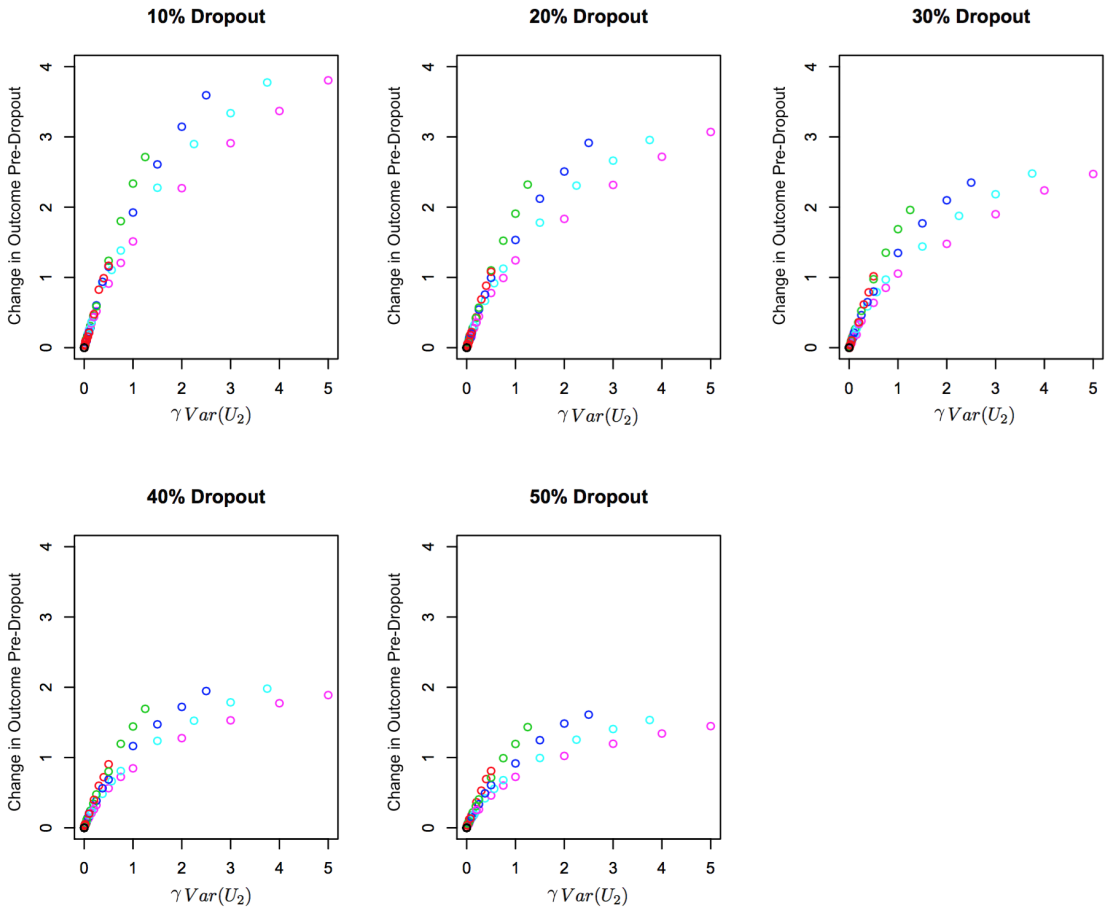


Figure 5.10: Plots of δy against $\gamma \times Var(U_2)$. For gammas; black = gamma 0, red = 0.1, green = 0.25, dark blue = 0.5, light blue = 0.75, purple = 1

The relationship between $\delta y'$ and $\gamma \times Var(U_2)$ can be clearly observed from Figure 5.10. The different colours highlight the different values of γ in the simulations. As $\gamma \times Var(U_2)$ increases, δy increases. However, this level of increase is proportional to the value of γ, and is smaller for higher γ values.

Model Fitting

As an attempt to quantify the relationship between δy and γ for this particular trial design, a mathematical formula is fit to the simulated data. A requirement for this formula is that

when $\gamma = 0$ the change in longitudinal outcome must be approximated as 0. Also the formula must be capable of taking into account $\gamma \times Var(U_2)$ and γ as separate outcomes when estimating δy.

A mathematical model of the form

$$\delta y = \hat{\omega}_1 \gamma + \hat{\omega}_2 \gamma Var(U_2) + \hat{\omega}_3 \gamma^2 Var(U_2) + \epsilon \tag{5.3}$$

was fit to the data using standard modelling techniques. Models were fit separately for each percentage of dropout where δy denotes the difference in longitudinal outcome between the last two time points. The following table shows the estimates of $\hat{\omega}_1$, $\hat{\omega}_2$ and $\hat{\omega}_3$ for each dropout percentage;

% Dropout	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$
10	0.4050	2.2752	-1.5558
20	0.3450	1.8738	-1.3093
30	0.3053	1.5984	-1.1490
40	0.2848	1.3703	-1.0318
50	0.2541	1.1701	-0.9231

Table 5.2: Mathematical Model Parameter Estimates

Diagnostic plots for the parameters of each of the above models are included in the appendix. From observing the q-q plots of each model, the residuals appear to be evenly distributed. A Shapiro-Wilk test of normality was employed and the values were significant at the 95% confidence level for each model above. The q-q plots appear to incite a successful model fit in all cases, particularly for lower percentages of dropout with the majority of points being on $y = x$. A higher number of outcome values are underestimated for higher percentages of dropout, with all simulation results highlighted in Table 5.1 being shown as underestimated in the q-q plot (in particular points 30 and 31 in the q-q plots of 40 and 50% correspond to the 2nd, 3rd, 7th and 8th rows in Table 5.1). The fact that these points were identified to be underestimating the change in longitudinal profile pre-dropout

within the model is a positive attribute.

In order to test the above models, data was simulated 1000 times for 500 patients per group for 5 different scenarios. One of the aims stated in Section 5.4 was to compare the results for flat baseline hazard simulations to alternative baseline hazard profiles. Therefore, three different variations of the simulations were carried out in order to test the appropriateness of the model. Data was simulated with a flat baseline hazard, an increasing baseline hazard and a decreasing baseline hazard drawn from the Gompertz distribution [113]. The mean longitudinal outcome change in the last two time points prior to dropout was recorded and the results compared to the approximated value from Equation 5.3. Table 5.3 shows these results;

% Dropout	γ	$Var(U_2)$	Pred. δy	Sim. δy (flat)	Sim. δy (inc)	Sim. δy (dec)
10	3	0.0002	1.2135	1.2129	1.2350	1.2006
20	0.3	0.0395	0.1211	0.1212	0.1254	0.1201
30	6	0.0002	1.8254	1.8221	1.8334	1.7991
40	0.07	6	0.5651	0.5623	0.5625	0.5611
50	1.2	0.8	0.7193	0.7190	0.7186	0.7179

Table 5.3: Model Verification

All the simulated values with a flat baseline were within 0.03 of the predicted change in longitudinal outcome pre-dropout, indicating a good model fit. However, the estimates were less accurate for a non flat baseline. It can be observed that when the baseline dropout risk was decreasing over time, the mean change in longitudinal outcome also decreased. The opposite was the case for an increasing baseline hazard, further demonstrating that the approximation for γ is also dependent on baseline properties. Despite this, the estimates were all within 0.03 of the correct dy for all verification simulations.

Classifying the γ parameter

One of the aims of this study was to provide a formula which could be used to observe the relationship between γ, δy, $Var(U_2)$ and percentage of dropout for the trial scenario in simulation study 2. The model established appears to provide a useful approximation to this relationship. In order to provide an approximate estimate for γ based on δy, Equation (5.3) can be rearranged.

As

$$\delta y = \hat{\omega}_1\gamma + \hat{\omega}_2\gamma Var(U_2) + \hat{\omega}_3\gamma^2 Var(U_2)$$

it follows that

$$0 = \hat{\omega}_1\gamma + \hat{\omega}_2\gamma Var(U_2) + \hat{\omega}_3\gamma^2 Var(U_2) - \delta y \tag{5.4}$$

$$0 = \hat{\omega}_3 Var(U_2)\gamma^2 + (\hat{\omega}_1 + \hat{\omega}_2 Var(U_2))\gamma - \delta y \tag{5.5}$$

Equation (5.5) has the following roots;

$$\gamma = \frac{-(\hat{\omega}_1 + \hat{\omega}_2 Var(U_2)) \pm \sqrt{(\hat{\omega}_1 + \hat{\omega}_2 Var(U_2))^2 + 4\hat{\omega}_3 Var(U_2)\delta y}}{2\hat{\omega}_3 Var(U_2)} \tag{5.6}$$

This leads to the following approximation formulae for each percentage dropout, with the smallest positive value being the estimate of γ due to the quadratic nature of the approximation and the difficulty in estimating δy for higher values of γ.

10% Dropout

$$\gamma = \frac{(0.4050 + 2.2752 \times Var(U_2)) \pm \sqrt{(0.405 + 2.2752 \times Var(U_2))^2 - 6.2232 \times Var(U_2)\delta y}}{3.1116 \times Var(U_2)} \tag{5.7}$$

20% Dropout

$$\gamma = \frac{(0.3450 + 1.8738 \times Var(U_2)) \pm \sqrt{(0.3450 + 1.8738 \times Var(U_2))^2 - 5.2372 \times Var(U_2)\delta y}}{2.6186 \times Var(U_2)} \tag{5.8}$$

30% Dropout

$$\gamma = \frac{(0.3053 + 1.5984 \times Var(U_2)) \pm \sqrt{(0.3053 + 1.5984 \times Var(U_2))^2 - 4.5960 \times Var(U_2)\delta y}}{2.2980 \times Var(U_2)} \tag{5.9}$$

40% Dropout

$$\gamma = \frac{(0.2848 + 1.3703 \times Var(U_2)) \pm \sqrt{(0.2848 + 1.3703 \times Var(U_2))^2 - 4.1272 \times Var(U_2)\delta y}}{2.0636 \times Var(U_2)} \tag{5.10}$$

50% Dropout

$$\gamma = \frac{(0.2541 + 1.1701 \times Var(U_2)) \pm \sqrt{(0.2541 + 1.1701 \times Var(U_2))^2 - 3.6924 \times Var(U_2)\delta y}}{1.8462 \times Var(U_2)} \tag{5.11}$$

It is important to note that the above formulae are not capable of approximating for all γ values, however the majority of plausible combinations are accounted for. Data has mainly been simulated in this chapter for trial scenarios where longitudinal readings are made at unit intervals over time. It is sometimes the case that there are larger gaps between equidistant timepoints, longer follow up times and different definitions of γ in joint modelling analyses. For these trials, the simulation study in this chapter should act as a guide for future γ estimation studies.

5.5 Conclusions

By using simulation methods, graphical representations of the relationship between γ and change in outcome pre-dropout are presented in this chapter. With the right level of information available this could provide useful to a clinician in estimating γ, as there are benefits to having knowledge about the properties of this value prior to the commencement of a study. From the results of the simulation study, it was established that the parameter estimate for γ is dependent on:

- The estimated percentage of dropout and baseline hazard
- The variance of the random slope U_2
- The change in longitudinal outcome pre-dropout.

The simulation study showed that the higher the values of γ and $Var(U_2)$, the greater the change in longitudinal outcome pre dropout, and that these two variables had multiplicative properties. Contrarily the change in longitudinal outcome was less prominent for higher percentages of dropout.

When considering the power and sample size of a longitudinal trial, the estimated percentage of dropout is usually taken into account during the planning phase [114]. Similarly, statistical methods exist for estimating the variance of the slope and intercept in a random effects model, as well as the possibility that details of such a variance can be estimated by using historical data in previously conducted trials. Using these factors it is possible to estimate γ using simulation methods as demonstrated in simulation study 2.

A clear example of how this can be done was demonstrated in simulation study 2. By simulating data sets for different dropout percentages and values of γ, $Var(U_1)$ and $Var(U_2)$ we were capable of establishing an approximate relationship between change in longitudinal outcome and γ estimate for the trial design described in Section 5.4. As the

design of a trial is generally discussed prior to the randomisation of patients, similar methods can be applied to any given trial design to approximate γ. As a guide, the R code based upon the joineR software [46] used for simulating data in the second simulation study is included in the Appendix.

Until now, there has been little discussion of diagnostic methods used in joint modelling, however some methods are available for this framework. In Chapter 6, joint modelling diagnostic procedures are introduced, and software is developed to carry out these methods in R.

Chapter 6

Diagnostics of Joint Modelling

6.1 Introduction

In Section 1.4, a history of the development of joint modelling was presented. Research in the area of joint modelling of longitudinal and time-to-event data is ongoing and many different specifications of the joint model have been proposed and implemented in published articles [49, 115, 116]. While a lot of progress has been made in this area, the focus of research has largely been based around model generation and procedures for parameter estimation. Therefore, there still are a large number of research questions which need to be addressed to develop a stronger and more complete framework in Joint Modelling [140].

The appropriateness of using joint modelling analyses are dependent on some untestable assumptions, however there are procedures available which can be used to assess the applicability of joint models being fit to a given dataset, and to identify outliers once a model has been fit [119,120]. Although many diagnostic procedures exist for separate longitudinal and time-to-event analyses [87,117,118], the development of diagnostics for joint modelling is still in the early stages, and few papers have been published which discuss the progress of diagnostics in this area [140].

There are two main articles concerning the development of joint modelling diagnostic procedures. Dobson and Henderson 2003 [119], proposed a selection of diagnostic methods for the joint models described in Henderson et al., whereas Rizopoulos 2007 focused on the area of residual analysis for joint models with a parametric time-to-event component [120]. Since the development of these diagnostic procedures, they have not been used in clinical trial analysis in published literature, and no statistical packages or software currently exist to carry out the methods proposed in the former paper. In this chapter, code in the statistical program R is developed to carry out a variation of the sequential discrimination method described in the original paper [119], and for the calculation of Cook's distance for the random slope and intercept joint model as specified in Section 1.4.3. These methods

are illustrated using the MAGNETIC analysis from Section 2.5.

As we have only discussed the semi-parametric Cox model in the thesis up until this point, in the final section of the chapter a relative risk model with Weibull baseline function is introduced with a view to illustrating the multiple imputation residual analysis procedure as described in Rizopoulos, 2007 [140]. Utilisation of the relative risk model and subsequent residual analysis will be demonstrated using the MAGNETIC dataset [47].

6.2 Sequential Discrimination

6.2.1 Methods

Graphs of Sequential Discrimination

As shown in the systematic review in Chapter 3, joint modelling has been a method rarely used in recent clinical trials [60]. While this may be due to joint modelling being a relatively newly developed procedure, fitting joint models is also more complicated than using standard linear mixed modelling techniques. By using sequential discrimination, it is possible to establish whether a longitudinal analysis alone is sufficient, or whether there are intrinsic differences between the patients that remained in the study and those that dropped out which fail to be accounted for by using separate analyses [119]. In these cases, joint modelling is a more appropriate method of analysis. Sequential discrimination provides both a visual demonstration and quantification of the prognostic differences between patients that dropped out and continued at each time point.

To distinguish a difference in longitudinal profiles between patients that left the study after a given time point and patients that continued, Dobson and Henderson propose a technique based on Diggle (1989) [122]. The method tests the difference in between dropouts and non dropouts at all longitudinal time points by initially categorising patients still active in the trial at each time point, d_j, into groups, ν_d, depending on their dropout

status at that time point. We define the risk set $\mathcal{R}_d$ to be all patients that could drop out at a given time point. For example, when testing for the differences between dropouts and non-dropouts at time $t = 20$, define $\nu_{20} = 1$ for each individual that had no more longitudinal readings after $t = 20$ and $\nu_{20} = 2$ for the patients that continued in the trial.

A standard linear mixed effects model, as described in Section 2.3.3, is then fit to the observed longitudinal data. In the original method described by Dobson et al., a linear model with a time parameter is fit to the residuals up until time point d_j for each individual still in the trial at this time. The parameter estimates for the residual slope are plotted against the parameter estimates of the residual intercept for each individual, and those that dropped out are distinguished from the non-dropouts.

However, in this thesis a variation of the original method is proposed in an attempt to establish a procedure which is both more understandable and easier to interpret for clinicians.

To carry out this variation of the original sequential discrimination, define

$$R_i = Y_i - x'_{1i}\beta_{\mathbf{1}} \tag{6.1}$$

as the vector of residuals for each subject, i, and classify those patients who are at risk of dropping out at each time point d into a risk set $\mathcal{H}_d$. A subset of the residuals $\mathbf{R_i}$ is then generated, denoted $\mathbf{R_{id}}$ which contains all the residuals for patients with a complete set of measurements prior to timepoint d.

We define the slope of the residuals as the difference between the residual at the given time point and the previous time point. A correlation matrix for the 2 dimensional summaries of the residuals and slope of residuals at each time point can be generated as these

summaries would be complete and balanced. For each separate time point, the slopes of the residuals are plotted against the residuals for patients in the risk set and separated into the pre-specified groups dependent on dropout profile.

By observing the graphs, any obvious differences between the patients that dropped out and stayed in at each time point can be identified by observing the distribution of residuals and change in residuals from the previous time point (denoted here as the slope of the residual). While the originally proposed method in Dobson and Henderson [119] is dependent on the historical profiles of patients, this newly proposed method focuses on the most recent longitudinal outcomes prior to dropout. This way any radical changes in outcome can be identified in the graphs and any profile differences in longitudinal outcome for patients that dropped out can be observed clearly. If there appears to be a clustering of patients in one group away from the other at various time points then this may indicate a difference in dropout profiles and that joint modelling methods may be reasonable to use.

Full details of the joint modelling analysis of MAGNETIC are provided in Chapter 2.5.2. To carry out the sequential discrimination for MAGNETIC, Equation 2.1 in Chapter 2 was fit to the data with $\beta_{\mathbf{1}} = \{\alpha, \beta_0, \beta_1\}$ where α is the parameter for time, β_0 is a fixed intercept and β_1 is the treatment effect. The residuals and slope of residuals were calculated for each time point.

In Section 6.2.2, two pre-joint modelling sequential assessments are carried out to establish the appropriateness of using joint modelling in the MAGNETIC trial using this newly proposed variation of the original method. The first aims to establish whether there were any intrinsic differences in the residual profiles between dropouts and non-dropouts at each time-point as described above. However, an additional analysis was also carried out to investigate the difference between the patients that dropped out due to good status, other dropouts and the patients that remained in the study using graphical methods. The

details of patients that dropped out due to good status was recorded by the trialist in MAGNETIC.

Quantifying Sequential Discrimination

While graphs can provide a visual overview of the differences in each group, it can sometimes be difficult to establish the extent of the general differences between dropouts and non-dropouts at each time point. To generate a p-value to test the difference in profiles between the groups, two-dimensional summaries of $\mathbf{R_{id}}$ denoted by $\mathbf{p_{id}} = (\nu_{id1}, \nu_{id2})$ are generated for each dropout group, with these ν_i referring to the aforementioned residuals and slope of the residuals. The differences between the two dimensional summaries for the different dropout groups at each timepoint are then tested using a discriminant analysis based on the Mahalanobis distance [123], defined as

$$M_{qr} = (\bar{\mathbf{p}}_{\mathbf{qd}} - \bar{\mathbf{p}}_{\mathbf{rd}})' \mathrm{Var}(\mathbf{p_{id}})^{-1} (\bar{\mathbf{p}}_{\mathbf{qd}} - \bar{\mathbf{p}}_{\mathbf{rd}}) \tag{6.2}$$

where $\bar{\mathbf{p}}_{\mathbf{gd}}$ is the sample mean of group g subjects. This generates a value for the Mahalanobis distance between the two groups for each dropout point. To test the significance of this Mahalanobis distance, p-values are generated by permuting the patients in each group, calculating the corresponding Mahalanobis distances for 10,000 different permutations of the group indicator and then comparing the value M_{jk} to these generated values.

For an analysis of the MAGNETIC data set, the Mahalanobis distance was calculated for the plots described in this section, and the p-values calculated. This was initially done to compare the dropouts to non-dropouts at each timepoint, and subsequently non-dropouts with patients that dropped out due to good status and other dropouts. Original R code developed for generation of the graphs, calculation of Mahalanobis distances and p-values is included in the appendix.

6.2.2 Results

An analysis was carried out to assess whether there was a difference in the profiles between dropouts and non dropouts in MAGNETIC. A linear mixed effects model was fit to the ASS score. The following are plots of the slopes of the residuals against the residuals for time points $t = 40$ onwards, due to the low number of dropouts before this time.

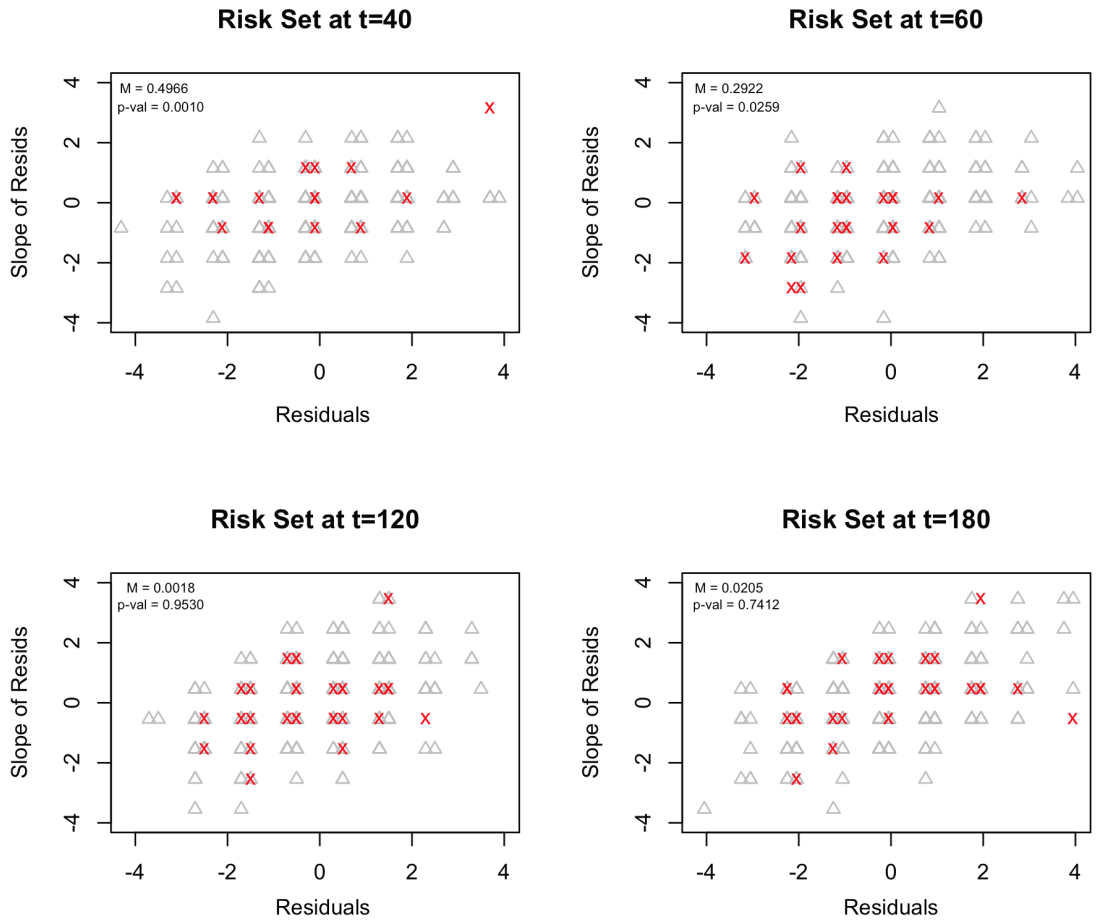


Figure 6.1: Sequential Plots showing patients that dropped out (red crosses) at each time point compared to those who did not (grey triangles). The values to the left of the graphs indicate the Mahalanobis distance and the p-value at each time point.

Due to the few fixed effect covariates included in the model and the discrete nature of the outcome, it was difficult to assess whether there were any intrinsically different charac-

teristics in the profiles of those patients that dropped out at each time point compared to those that did not simply by observing the graphs. Higher values on the y-axis indicates an increase in the residual from the previous time point, while higher x-axis values indicate a higher than expected longitudinal outcome. At $t = 40$, a large proportion of patients that dropped out had negative values for the slope of residuals, although the individual with the highest residual profile also dropped out at this timepoint. At $t = 60$, it would appear that there were a higher proportion dropouts with a negative residual than non-dropouts, and therefore lower longitudinal values than expected. The range of the slope residuals was higher for patients that remained in the study, with the majority of slope values ranging between -2 and 2 for patients that dropped out at each time point.

When observing the p-value of the Mahalanobis distance, it can be seen that there is a significant difference in the profiles of the two groups at $t = 40$ and $t = 60$. The Mahalanobis distance at $t = 120$ and $t = 180$ was not found to be significant. It is interesting to note that patients stopped receiving treatment at $t = 60$, after which the longitudinal profiles showed no significant difference for dropouts compared to continuers in the trial. These results suggests that a joint modelling analysis may be appropriate, as the Mahalanobis distance is significant at some time points.

In the analysis of the MAGNETIC trial, the results (as listed in Section 2.5.2, Table 2.2) suggested that patients with were more likely to dropout due to good status. Details of some of the reasons for dropout were collected by trialists, of which some patients were dismissed due to good status. However, other patients that left the study had no recorded reasons for dropout. We now compare the profiles of non-dropouts at each time point to both those that dropped out due to good status and other dropouts. Figure 6.2 shows the of slope of residuals plotted against the residuals.

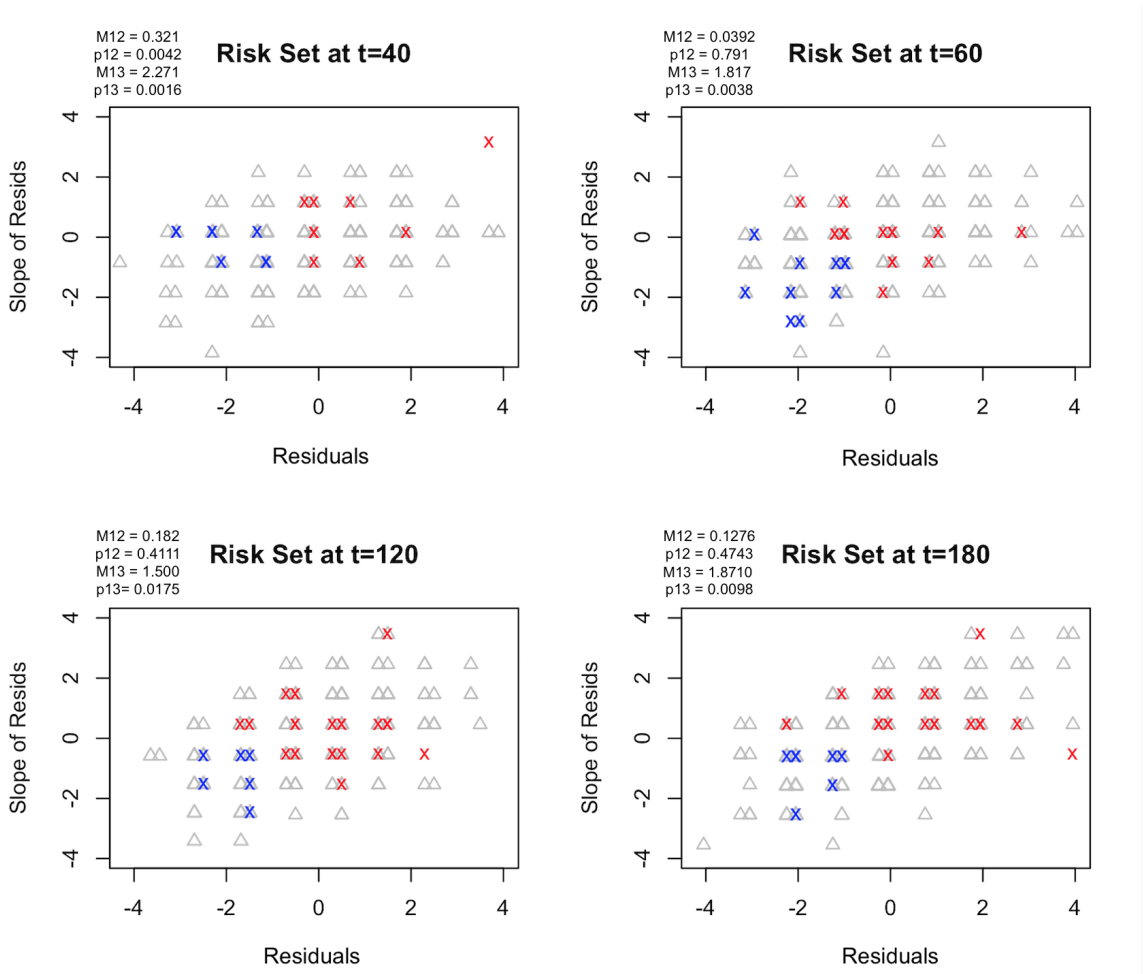


Figure 6.2: Sequential Plots showing patients that dropped out non-informatively (red crosses) at each time point compared to those who dropped out due to low severity score (blue crosses) and those who did not dropout(grey triangles). The values to the left of the graphs indicate the Mahalanobis distance and the p-value at each time point between groups. Group 1 = Non dropouts, Group 2 = Dropouts not due to good status, Group 3 = Dropouts due to good status.

The graphical differences in the residual properties of those patients that dropped out due to negative status compared to the other patients in the study is clearly observed from Figure 6.2. Analysing this quantitively, it was found that there is a highly significant difference in the profiles of those that dropped out due to good status at all time points when compared to non-dropouts. There was only found to be a significant difference

between patients that dropped out for alternative reasons and continuers in the trial at $t = 40$.

6.3 Case Influence

6.3.1 Methods

Once the model has been fit, the influence different subjects have on the regression parameters $\theta = (\beta_{\mathbf{1}}, \beta_{\mathbf{2}}, \gamma)$ can be analysed. The standard method of calculating this is using the Cook's distance [130] . Implementation of this method can be more complicated in joint modelling procedures than standard estimation due to additional parameters required for estimation. To calculate the Cook's distances, patient i is excluded from the analysis, the model is refit and the parameters without individual i, $\hat{\theta}_{(i)}$ are calculated for each subject [131]. These can then be substituted into the following formula which calculates the Cook's Distance for patient i:

$$CD_i = (\hat{\theta}_{(i)} - \hat{\theta})' Var(\hat{\theta})(\hat{\theta}_{(i)} - \hat{\theta}) \tag{6.3}$$

Original code for calculating the Cook's Distances of a joint modelling data set has been included in the appendix. It has been suggested in the past, that calculating the correct Cook's Distance can be impractical and time consuming [119]. However, due to the increase in computing power and efficiency, this is no longer the case.

However, to test an alternative approach, we also propose an approximation to Cook's Distance as proposed in Dobson et al. 2003:

$$\hat{\theta}_{(i)} - \hat{\theta} = -I^{-1}_{(i)}(\hat{\theta}) U_{(i)}(\hat{\theta}) \tag{6.4}$$

which utilised the negative of the observed score multiplied by the information. The success of this approach will be tested in this chapter. Code for calculating the observed score

and information is included in the JM package [121].

The methods of case influence described above are applied to the MAGNETIC data set. The aims of this are to firstly identify any outliers in the MAGNETIC analysis for further investigation and secondly to compare the results of using the true Cook's distance for each patient and the approximated Cook's distance using the estimation in equation 6.4. After identifying the most influential patients, the longitudinal profiles of these will be plotted and discussed. While Cook's distance can be used to focus on specific parameters, here we apply it to the full model fit and therefore define $\theta = \{\alpha, \beta_0, \beta_1, \beta_2, \gamma\}$. Outliers are more frequently identified by observation rather than application of a mathematical definition, due to the diverse range of potential specifications of an outlier. In this study, we define outliers as any patients with a Cook's distance that is both considerably higher than both the majority of patients, and greater than 3 times the mean of the Cook's distances.

6.3.2 Results

After the model was fit, a case influence was carried out based on the Cook's distance to highlight any influential patients. Figure 6.4 shows the true Cook's distance for the 508 patients.

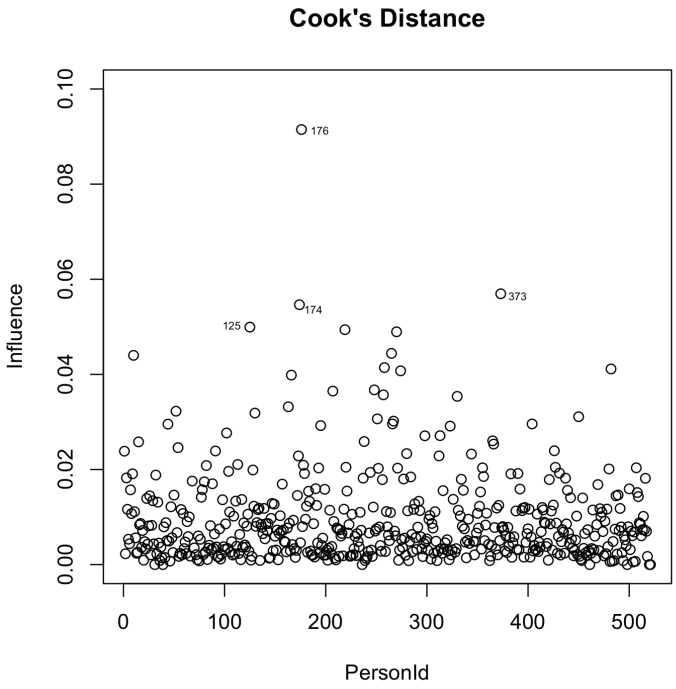


Figure 6.3: Cook's Distance for each patient

It can be observed from the graph that the most influential patients were those with ID's equal to 176, 373, 174 and 125. Patient 176 has a considerably greater influence on parameter estimates than anyone else, and this was therefore identified as an outlier. Figure 6.4 shows the longitudinal responses for the four most influential patients.

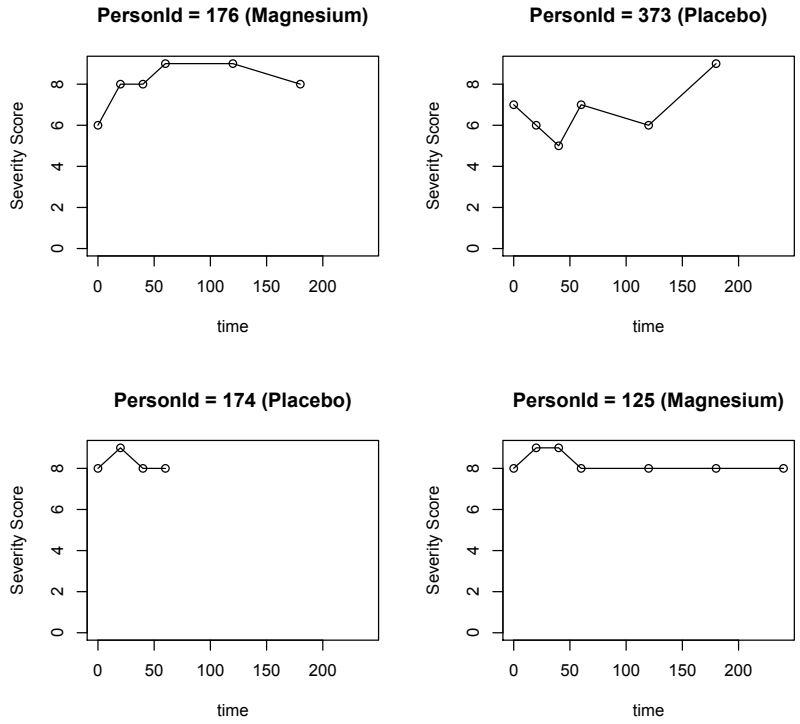


Figure 6.4: Plots showing the severity profiles of the most influential patients

The graph shows that the most influential patient was randomised to Magnesium. Unlike the majority of patients, the most influential individual (176) had an initial increase in severity score. In the first 60 minutes since randomisation, the ASS for this patient increased from 6 to 9, the maximum severity, indicating a deterioration in the condition. While the majority of patients given Magnesium recorded lower ASS scores, this patient had the maximum recorded on two separate occasions. Also, the γ estimate indicated that patients with a lower ASS were more likely to dropout, however patient 176 dropped out from the study with an ASS of 8, which is categorised as severe.

Patient 373, randomised to placebo, initially showed improvements in their condition. However, their ASS increased from 6 to 9 within the last 60 minutes of follow up before they dropped out from the study due to poor status. This high score at the time of drop

out may have been the reason for the high influence of this patient. Patient 174 was also randomised to placebo, and entered the study with a high severity of asthma. After 60 minutes of treatment, this patient dropped out without any indication of an improvement in ASS. Unlike the other influential cases, patient 125 did not dropout, however their ASS was consistently high despite being randomised to magnesium.

Analysis of case influence using the true Cook's distances was then compared to the results of the approximation, to test the effectiveness. The results are shown in Figure 6.5

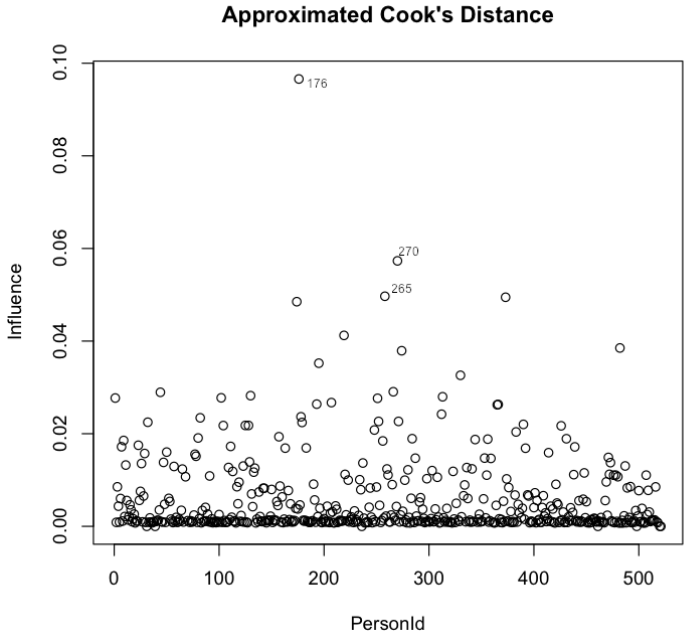


Figure 6.5: Approximate Cook's Distance using First Order Estimate

The Pearson's correlation coefficient between the correct Cook's distance and the approximate Cook's distance was found to be 0.79. However, while patient 176 was identified as the most influential, the 2nd and 3rd most influential patients identified above were different to those in Figure 6.4. Therefore, while the approximation can provide similar results to the actual Cook's distance, it can be observed that the patients identified as influential are not always consistent across the two methods.

6.4 Residual Analysis using Multiple Imputation Methods

6.4.1 Methods

When carrying out separate longitudinal and time-to-event analyses of a data set, there are a multitude of available methods for residual analysis [87, 117, 118]. In a longitudinal context with a linear mixed model, the subject specific residuals or marginal residuals can be calculated to assess the model fit.

These can be used to carry out a residual analysis based solely on the observed longitudinal data. However, there may be assumptions within the standard definitions of these residuals that distort the accuracy of the model verification [140]. In particular, joint modelling is reliant on the assumption that the occurrence of dropout is potentially related to the subject-specific longitudinal profiles, and is therefore non-random. Therefore, residual analysis from the observed data alone may not be applicable, as the residuals may not display the typically associated properties (i.e. independence, mean zero etc.) [120].

To account for some of these issues caused by the nature of a non random dropout structure within a dataset for a residual analysis of the joint modelling longitudinal process, multiple imputation methods can be utilised [120, 126, 127]. However, this method requires a complete likelihood specification, and therefore has only been described in the literature as a method to be used for joint models with a parametric time-to-event component. While this method may therefore not be applicable to the main model utilised throughout this thesis, it is important to include details of this residual analysis as a) It is useful to demonstrate how alternative joint modelling specifications can be applied to a dataset, b) a demonstration may be beneficial for those who wish to fit a joint model with a parametric time-to-event component and c) to discuss the issues associated with merely carrying out a residual analysis based on the observed data only.

The aim of this method is to impute longitudinal data under the assumptions of the complete data model. These imputations are designed to replicate what may have happened if each patient had not dropped out, and therefore induce a complete dataset. Software for this type of modelling and residual analysis has been developed in the JM package in R [121].

To demonstrate this residual analysis, we fit a relative risk model with Weibull baseline hazard. The longitudinal component of the model is as given in Equation (1.1) from Chapter 1. The time-to-event element is modelled using the form:

$$\lambda(t) = \lambda_0(t) \exp\{x'_{2i}\beta_{\mathbf{2}} + \gamma m_i(t)\} \tag{6.5}$$

where $m_i(t)$ is the true unobserved longitudinal trajectory obtained by excluding the measurement error, $\beta_{\mathbf{2}} = \{\upsilon, \beta_2\}$ where υ is an intercept for the time-to-event element of the model, and γ is the association parameter. The baseline function is modelling by

$$\lambda_0(t) = \sigma_t t^{\sigma_t - 1} \tag{6.6}$$

where σ_t is the shape parameter. The residuals can be defined by

$$R_i(t_{ij}) = \{y_i(t_{ij}) - x'_{1i}\beta_{\mathbf{1}}(t_{ij}) - W_{1i}\}/\hat{\sigma}_y \tag{6.7}$$

To carry out these imputations, joint modelling must be considered in a Bayesian context. Samples are taken from the posterior distribution of the longitudinal outcome y^{mis}, where mis refers to the missing data, which is averaged over the posterior of the other parameters. The density of the distribution can be expressed as

$$p(y_i^{mis}|y_i^{obs}, T_i, d_i) = \int p(y_i^{mis}|y_i^{obs}, T_i, d_i; \theta) p(\theta|y_i^{obs}, T_i, d_i) d\theta_i \tag{6.8}$$

where y_i^{obs} is the observed data for a patient, T_i is the time of observation, and d_i is the dropout time. Let $U_i = \{U_{1i}, U_{2i}\}$ be the set of random effects for each patient i. Rizopoulos [140] showed that

$$p(y_i^{mis}|y_i^{obs}, T_i, d_i; \theta) = \int p(y_i^{mis}|U_i; \theta) p(U_i|y_i^{obs}, T_i, d_i; \theta) d\theta_i \tag{6.9}$$

and Cox and Hinckley [133] showed that $\{\theta, y_i^{obs}|, T_i, d_i\}$ can be approximated using $\theta \sim \mathcal{N}(\hat{\theta}, \text{Var}(\hat{\theta}))$. Therefore, the following procedure was proposed to generate the missing data;

- For imputation number $l = 1, ..., L$ draw the fixed parameters $\theta^{(l)} \sim \mathcal{N}(\hat{\theta}, \text{Var}(\hat{\theta}))$
- Draw random effects $U_i^l \sim \{U_i|y_i^{obs}, T_i, d_i, \theta^{(l)}\}$
- Generate the imputed values by drawing $y_i^{mis(l)}(t_{ij}) \sim \mathcal{N}\{\hat{m}_i^{(l)}(t_{ij}), \hat{\sigma_y}^{2,(l)}\}$.

In the first and last steps, samples are taken from the multivariate normal distribution. Step 2 in the above algorithm can prove to be computationally difficult as it is based on the posterior distribution of the random effects, and is in a non specified form. Therefore the Metropolis-Hastings algorithm, based upon proposals from the multivariate t distribution centred around Bayes estimates of U_i with a scale matrix $\text{Var}(U_i)$, is employed to carry out this sampling method. [120, 128, 140]

Once calculated, the value of the residuals are plotted against the fitted values and the LOESS smoothing [129] is used to plot a curve for the residuals of the observed data and the potential full data set with multiply imputed values. If the smoothed line is approximately equal to 0 for most fitted values, then this suggests a good fit. These residuals will be calculated for the model fit to MAGNETIC, as described in this section. Code for these procedures can also be found in the JM package for R [121].

6.4.2 Results

The Weibull relative risk model was fit to the ASS score in MAGNETIC, with the results of the model fit shown in Table 6.1.

Component	Parameter	Estimate	SE	95%Lower	95%Upper
Longitudinal	(Intercept)	5.6281	0.0743	5.483	5.775
	time	-0.0077	0.0003	-0.0083	-0.0070
	β_1	-0.2115	0.0489	-0.4069	-0.0161
Survival	β_2	0.5202	0.1881	0.1515	0.8889
Association	γ	-0.0920	0.0884	-0.2652	0.0812

Table 6.1: Results of joint model fit to ASS for Weibull baseline

With this alternative model fit, the ASS score in the magnesium group was found to be significantly lower than in the placebo group. Furthermore, the results show that patients in the magnesium group were significantly more likely to dropout, so fitting the relative risk model with Weibull baseline yielded the same conclusions of treatment effect as the random slope and intercept joint model (Section 2.5). However, based on this different definition of γ, the association parameter was found not to be significant.

To check the success of the joint model, marginal residuals were plotted against the fitted values. To obtain an accurate estimation of the residual values, 50 imputations were used, which is standard in joint modelling literature [140]. Two LOESS smoothed curves are fit to the plot to assess the mean residual profiles. This was fit for both the observed values only, and the observed coupled with the multiply imputed values. The results are displayed in Figure 6.3.

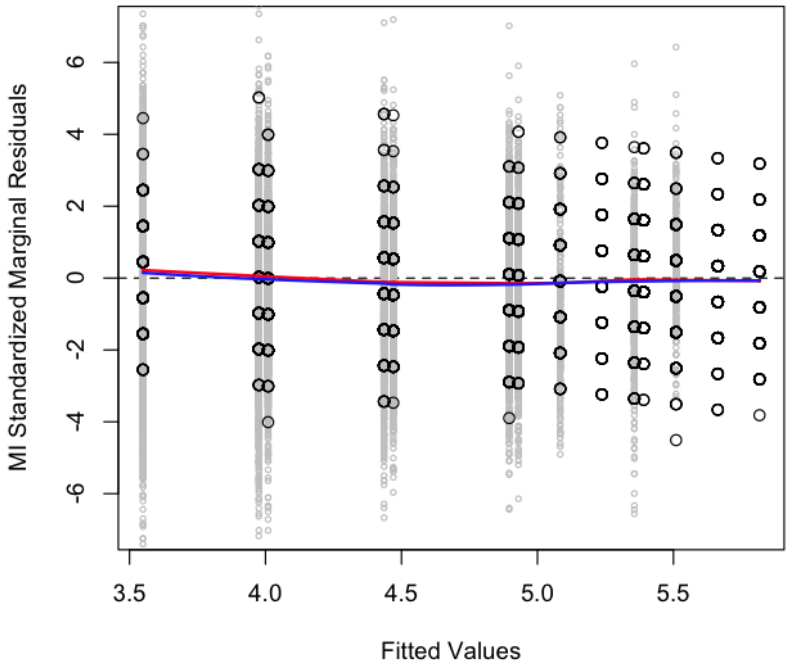


Figure 6.6: Plot of residuals against fitted values. The black circles indicate the observed residuals, with the grey points showing the multiply imputed values. LOESS smooth function for observed residuals (red), LOESS smooth function for both observed and imputed residuals (blue).

While the functional values at $y = 3.55$, are greater than 0, which decreases to below 0 at around $y = 4.45$, there is no large variation in the mean standardised marginal residuals. Observing the graph, the LOESS smooth functions for the observed values and the full imputed data set both maintain a trajectory close to $y = 0$. This indicates a good model fit for the joint model of MAGNETIC data.

6.5 Discussion

Due to the time constraints within a clinical trial, it is necessary to ensure that the most efficient and accurate method of analysis is carried out, and overcomplicated or inappropri-

ate analyses can lead to time being wasted. While the benefits of joint modelling have been highlighted throughout this thesis, this method requires a greater level of computational power than some standard methods which focus solely on longitudinal data. Sequential discrimination is used to check whether differences in profiles exist between dropouts and non-dropouts, which may indicate the need for joint modelling. The newly updated variation of sequential discrimination proposed in this thesis serves the same purpose as the original method, while simultaneously allowing for an easier interpretation of the plots for clinicians; as the emphasis is on the longitudinal outcomes immediately prior to dropout. While the original code generated for the diagnostic methods and included in the appendix is specific to MAGNETIC trial, more user-friendly functions for all the above will be included in the next release of joineR for R [46]. In the MAGNETIC data set, there was found to be a significant difference in profiles between dropouts and non dropouts at the earlier time-points in the study. Therefore a longitudinal analysis alone may not have accounted for the dropouts in a reasonable way, so it was appropriate to use a joint modelling analysis. This seems to concur with the analysis of ASS score in Section 2.5.2, which showed a significant γ estimate. One drawback to be considered within the analyses in this chapter, is that the ASS score is a discrete outcome, and therefore the differences between dropouts and non-dropouts were difficult to observe using graphical methods alone for sequential discrimination (see Figure 6.1). In a trial with continuous scores being taken, clusters of patients within the different dropout groups would be more evident.

To establish the existence of outliers and identify influential patients in a study using joint models, the Cook's distance for each patient is calculated. In the MAGNETIC trial, the 4 most influential patients were identified, all of which had considerably higher longitudinal profiles for ASS than the majority. However, the same patients were not identified as most influential when using an approximation to the Cook's distance. Historically, calculating Cook's distance was a lengthy procedure, however this is no longer the case for the majority of datasets. The code in the appendix demonstrates how this was approached for

the MAGNETIC trial. A more general original function capable of calculating the Cook's distance for other datasets and trial designs will also be included in the next release of joineR [46].

Standard methods of residual analysis may not account for dropouts effectively when applied to joint models, so a residual analysis based on multiple imputation methods was proposed for joint models with a parametric time-to-event component. This was used to check a relative risk Weibull model fit to the MAGNETIC dataset. The results from the residual analysis appeared to indicate a good model fit. While this method may also be capable of assessing model fit for multiple outcomes, no current method of residual analysis is available for the model proposed in Section 1.4.3 which is based upon the complete data.

In this thesis, methods have been described for joint modelling analysis, formulae for sample size estimation has been derived and code has been generated for diagnostic methods. However, these have mainly been applied to the MAGNETIC ASS analysis or simulations based upon MAGNETIC. The next chapter demonstrates applicability of joint modelling to a wider range of studies.

Chapter 7

Applications

7.1 Introduction

In Chapter 3, a systematic review highlighted that in longitudinal trials, joint modelling is used rarely in practice [60]. In some studies, having additional information about the relationship between an outcome and dropout or an event can be informative when evaluating drug effectiveness [134] - [138]. For example, in trials where a significant number of patients have a deterioration in outcome pre-event, further investigations can be carried out to evaluate the reasons for this pattern. Joint modelling is a feasible option for these studies. [140]

So why is joint modelling not used more frequently? This could be for a number of reasons. Firstly joint modelling is a relatively new methodology with the greatest advances in model development having been made in the past 15 years, and therefore the theoretical elements are still being developed. Consequently many trialists may be unaware of the methods of joint modelling. With sample size formulae usually reliant on simulations [140] and the majority of the methods being described in Chapter 6 not having been applied in practice, there are few examples of how joint modelling can be utilised for clinical trial data. In this Chapter, the methods described and developed in this thesis are applied to other clinical trial data sets. This is with a view to demonstrating the flexibility of the methodology and newly written software, providing examples of how joint modelling methods can be formally applied in RCTs. A secondary aim of this chapter is to gain a greater insight into the results of the MAGNETIC trial by analysing the data of the subcomponents [37, 47].

In this Chapter, the joint modelling methods outlined in this thesis will be used to analyse the subcomponents of the MAGNETIC trial, a trial for patients with minor mental health problems [46] and a trial involving treatment for patients with liver cirrhosis [56].

7.2 The Subcomponents Of ASS In The MAGNETIC trial

The methods described in this thesis have been demonstrated by being applied to the ASS outcome of the MAGNETIC trial [47], which is described in detail in Section 1.3. The conclusion drawn from investigating the properties of the ASS score is that joint modelling was an appropriate method of analysis for this dataset (See Section 2.4, Section 6.2). The model fit in Chapter 2 showed that ASS was significantly lower in the magnesium treatment group, patients randomised to magnesium were more likely to drop out and patients with lower severity scores dropped out more frequently in this trial. ASS is made up of three components measured at each time point; wheeze score, heart rate and muscle score rated from 0-3. Descriptions of the criteria for each can be found in Yung et al (1996) [38].

By analysing the subcomponents data, it may be possible to establish whether magnesium had a direct impact on one of the subcomponents more than the others and therefore obtain an understanding for the pharmacodynamic effect of magnesium on the body when treating acute severe asthma [139]. The analysis in Section 2.4 also demonstrated that patients with an improving prognosis were more likely to stop receiving treatment. By estimating γ for the subcomponents in MAGNETIC, we can hypothesise about which indicators may have been used by patients and healthcare professional to establish whether individuals had recovered.

Prior to a joint modelling analysis, sequential discrimination as described in Chapter 6.2.1 is applied to determine whether joint models are required for each subcomponent by initially fitting a linear mixed model for the longitudinal part. Mahalanobis distance [123] is calculated based on two dimensional summaries of the residuals and slope of the residuals for each time point, and the p-values are generated. In the cases where joint modelling may be appropriate, the model described by Equations (1.1) and (1.2) in Chapter 1 are fit to wheeze score, heart rate and muscle score data separately. For this analysis, we define

$\beta_{\mathbf{1}} = \{\alpha, \beta_0, \beta_1\}$ and $\beta_{\mathbf{2}} = \{\beta_2\}$. After model fitting, Cook's Distances are calculated to identify outliers [119], and the outcome profiles of these patients are examined.

Wheeze Score

For this analysis, the wheeze score data was treated as independent of the overall severity score analysis. This meant that as some patients had wheeze score available at some time points but not severity score, time-to-dropout was defined as the first time point at which a patient was missing a wheeze score measurement. Overall, 17.13% of children had at least one missing value for wheeze score, which was measured at baseline and 20, 40, 60, 120, 180 and 240 minutes post randomisation. Figure 7.1 shows the mean dropout profiles for patients randomised to Magnesium and Placebo.

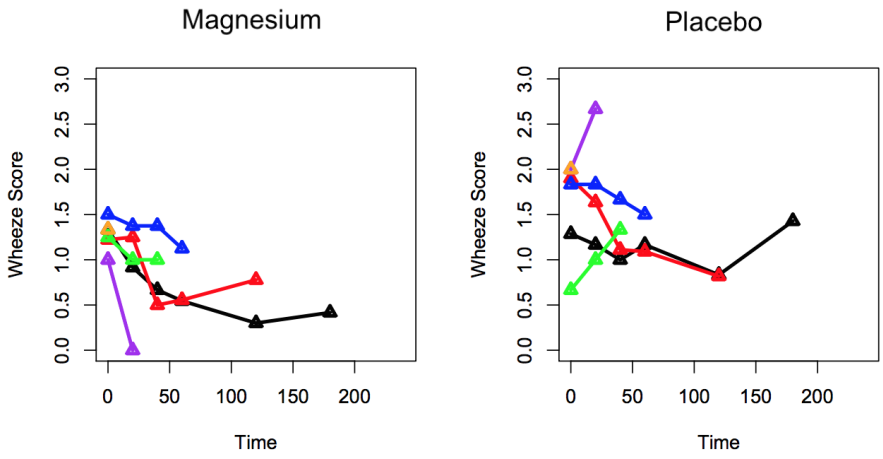


Figure 7.1: Mean Dropout Profiles for Wheeze Score are each time point

In the Magnesium group, the mean profiles of wheeze score for the patients that dropped out before $t = 60$ was either decreasing or constant. This is in contrast to patients that dropped out after $t = 120$, which displayed increasing mean profiles. In the placebo group,

the patients who had their final measurement at $t = 20$, $t = 40$ and $t = 180$ had increasing mean profiles. At all other time points, there was a decrease pre-dropout. The overall mean change in outcome pre-dropout was found to be negative.

The diverse range of mean profiles caused difficulty in establishing if a joint modelling analysis is appropriate for wheeze score. Therefore, sequential discrimination is used to check any differences in profile between dropouts and non-dropouts. Figure 7.2 shows the plots of sequential discrimination for times $t = 40, 60, 120, 180$. A plot for $t = 20$ is not included as there too were few dropouts to detect an informative difference in profiles.

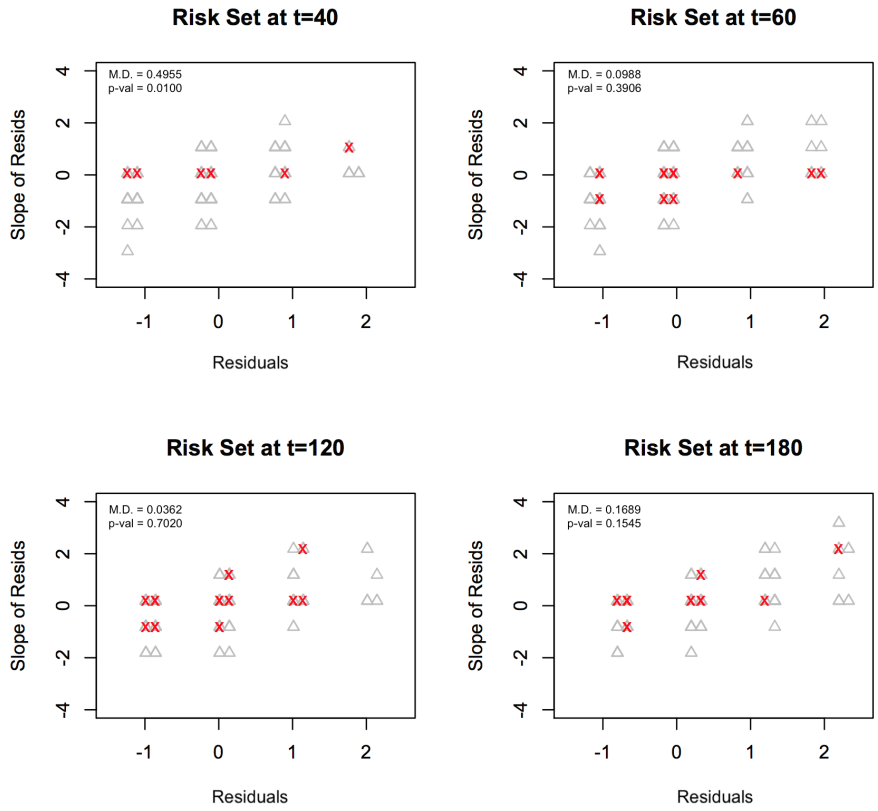


Figure 7.2: Sequential Discrimination for non-dropouts (grey triangles) v. dropouts (red crosses) at each time point

Figure 7.2 demonstrates that the range of values of the residual slopes for dropouts

was smaller than patients who remained in the trial. An analysis of Mahalanobis distance showed no significant difference in profiles at $t = 60, 120$ and 180. However, the low p-value at $t = 40$ suggests that early stage dropouts had different profiles to continuers. Therefore, mixed linear modelling techniques alone may not be sufficient to accurately model the longitudinal outcome, and joint models are fit to the data. The results of model fit are shown in Table 7.1.

Component	Parameter	Estimate	SE	95%Lower	95%Upper	P-Value
Longitudinal	(Intercept)	1.365	0.0368	1.291	1.424	
	time	-0.0031	1.7×10^{-4}	-0.0035	-0.0028	< 0.001
	β_1	-0.1370	0.0489	-0.216	-0.0354	0.005
Survival[6]	β_2	0.628	0.208	0.250	1.063	0.003
Association	γ	-0.255	0.283	-0.996	0.271	0.368
Variance	U_1	-0.332	0.0249	0.275	0.371	
	U_2	7.4×10^{-6}	1.6×10^{-6}	4.6×10^{-6}	9.0×10^{-6}	

Table 7.1: Results of joint model fit to wheeze score for MAGNETIC

A joint model fit determined that children randomised to magnesium had on average a wheeze score of 0.1370 lower than the placebo group, with 95% certainty that this value was between 0.035 and 0.216. Therefore this was found to be a significant difference (p=0.005). Also patients randomised to magnesium were more likely to dropout, with the 95% confidence interval for the log-hazard ratio of [0.250,1.063]. The parameter estimate for γ indicates that a lower score may be linked to dropout, however this result was not significant (p=0.368). In order to assess the presence of outliers, the Cook's distances are calculated to identify any patients with a high influence. This plot is shown in Figure 7.3.

[6] β_2 is reported on the log-hazard scale throughout the chapter

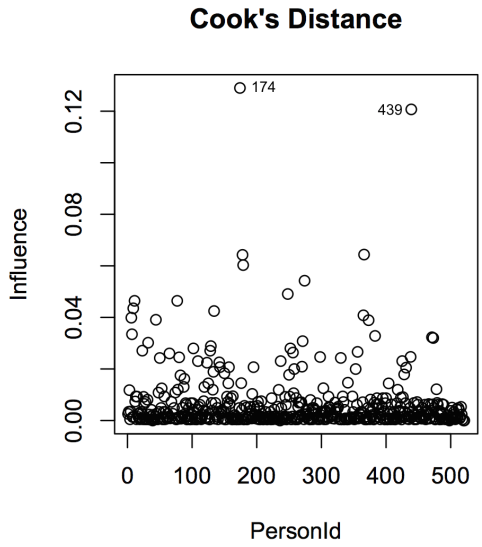


Figure 7.3: Plot of Influence for Each Patient

From Figure 7.3, patients 174 and 439 were identified as highly influential and we judged these to be outliers. Patient 174 was also identified as one of the most influential patients when modelling the overall ASS score. Figure 7.4 shows the longitudinal wheeze score profiles for the outliers in this model.

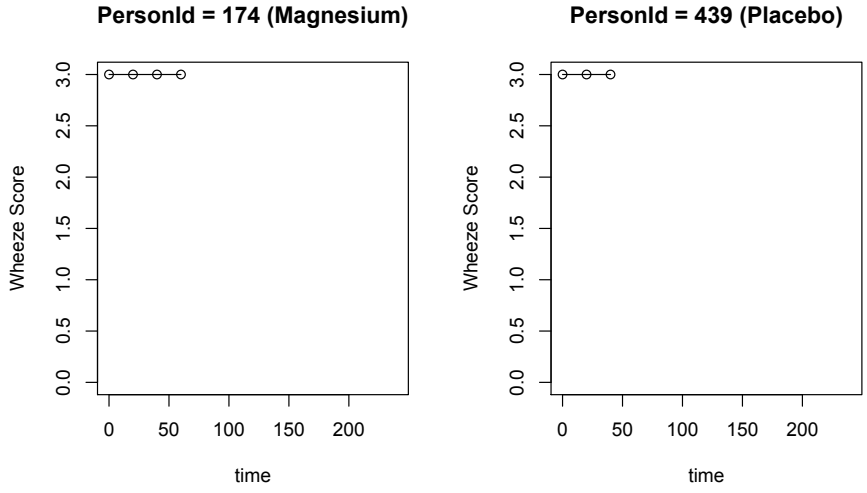


Figure 7.4: Longitudinal Profiles of Most Influential Cases

Both influential patients had the maximum wheeze score prior to dropout, which was considerably higher than average. While in general the wheeze score over time was decreasing for the dataset, no change in outcome was observed prior to dropout for these patients. The negative estimate for γ suggests that patients with lower outcomes were more likely to drop out, which was not the case for the outlying patients.

Heart Score

In the following section, the heart score data for MAGNETIC is analysed. For the heart score subcomponent, 18.5% of patients had at least one missing value. The mean dropout profiles for heart score are shown in Figure 7.5.

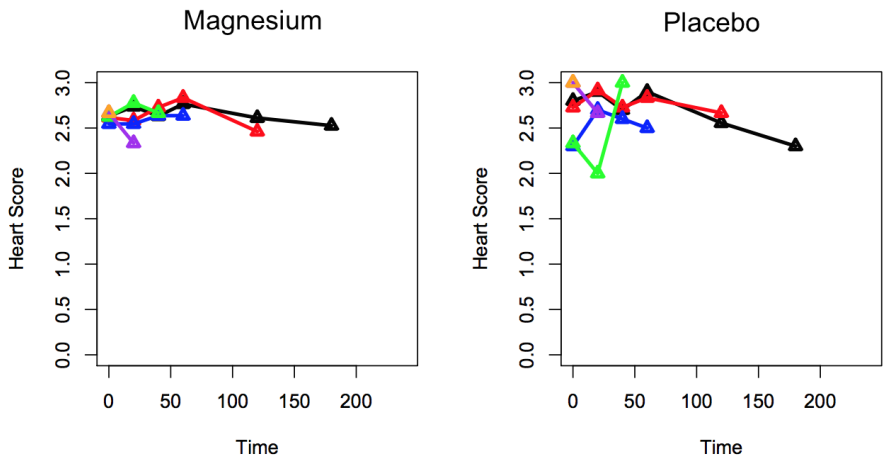


Figure 7.5: Mean Dropout Profiles for Heart Score are each time point

The group of patients that dropped out at $t = 60$ had a mean increasing heart rate in the magnesium treatment group. However, the mean dropout profile at the other time points had a decreasing score prior to the time of withdrawal. The average dropout profile in the placebo group for patients at $t = 40$ show a comparatively large increase in heart score prior

to withdrawal. Otherwise the mean profiles were decreasing. In general, heart scores were found to be higher than wheeze and muscle use scores, and there was less variation in the dropout profiles for heart score. To investigate the need for joint modelling further, Figure 7.6 shows the sequential discrimination at each time point $= 40, 60, 120, 180$, calculated by fitting a linear mixed model.

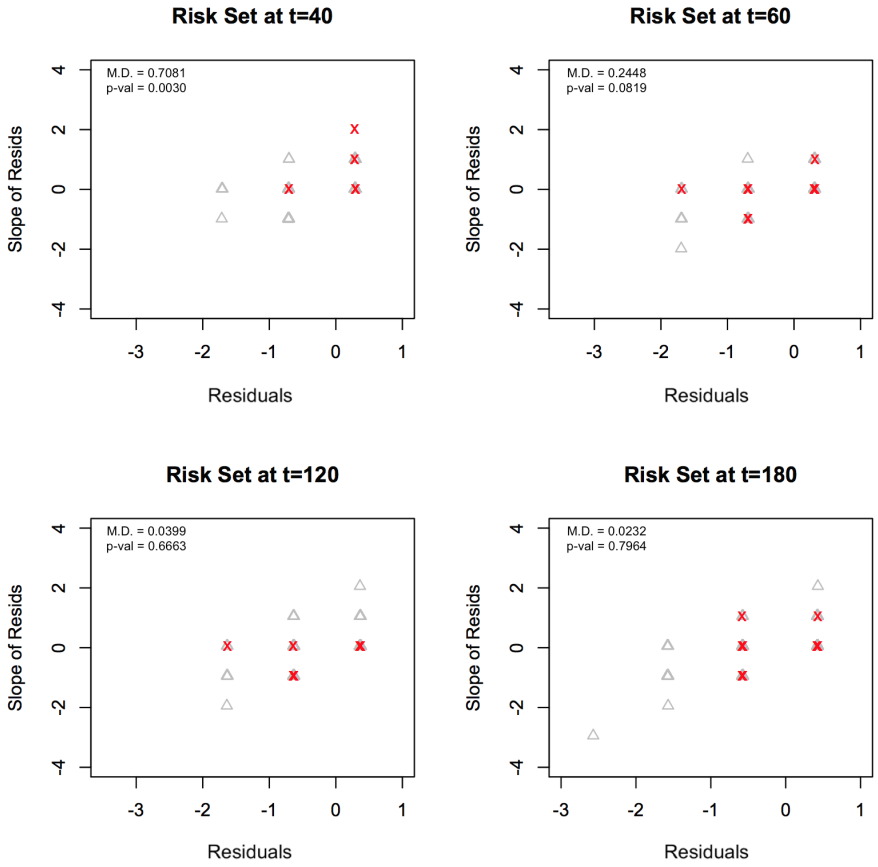


Figure 7.6: Sequential Discrimination for non-dropouts (grey triangles) v. dropouts (red crosses) at each time point

Overall the variation of residuals and residual slopes for heart score was less spread out than for wheeze score. The differences in profiles at $t = 60$, $t = 120$ and $t = 180$ was not found to be significant, despite a larger Mahalanobis distance at $t = 60$. However, the results of the sequential discrimination indicated a significant difference in residual profiles

for dropouts at $t = 40$, so joint modelling may be necessary for analysing the heart score data in MAGNETIC. A joint model is fitted to estimate γ, β_1 and β_2, and the results are detailed in Table 7.2.

Component	Parameter	Estimate	SE	95%Lower	95%Upper	P-Value
Longitudinal	(Intercept)	2.751	0.0262	2.684	2.786	
	time	-0.0010	1.3×10^{-4}	-0.0013	-8×10^{-4}	< 0.001
	β_1	-0.0102	0.0361	-0.0798	0.0634	0.778
Survival	β_2	0.456	0.218	0.0402	1.028	0.033
Association	γ	-0.394	0.236	-0.850	0.0438	0.095
Variance	U_1	0.124	0.0144	-0.0943	0.154	
	U_2	3.1×10^6	5.0×10^{-7}	2.4×10^{-6}	4.1×10^{-6}	

Table 7.2: Results of joint model fit to heart score for MAGNETIC

By analysing the joint model applied to heart score, it is observed that the longitudinal difference in heart score between the two groups is not found to be significant (p=0.778). The estimate for β_1 is -0.0102. This represents approximately one hundredth of a heart score point, which is clinically negligible. Patients with lower heart scores were more likely to drop out in this study although this was not significant, as indicated by the γ parameter (p=0.095), with a confidence interval of [-0.850,0.044]. The β_2 parameter was found to be significant in this study (p=0.033). Figure 7.7 shows the Cook's distances for each patient.

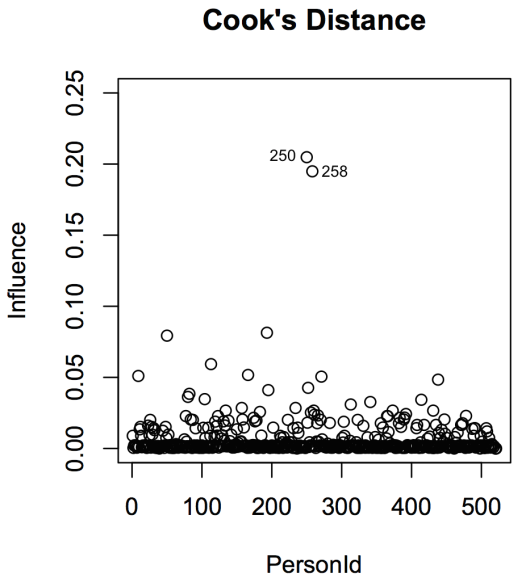


Figure 7.7: Plot of Influence for Each Patient

From the influence diagnostic, we classified patient's 250 and 258 as outliers and the most influential values. Longitudinal profiles for these patients are presented in Figure 7.8.

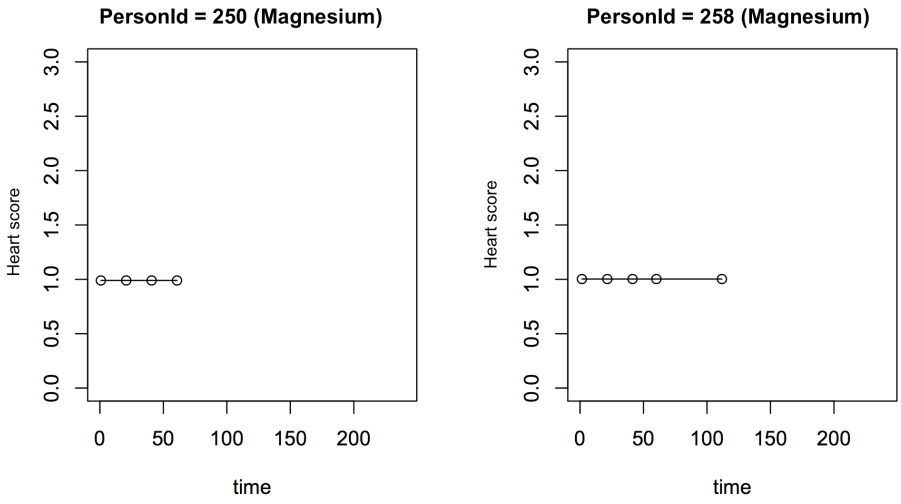


Figure 7.8: Longitudinal Profiles of Most Influential Cases

Both influential patients had a constant heart score of 1 over time, despite the estimate of the intercept for the heart score being 2.751. Both patients were randomised to magnesium, which on average had patients with lower heart scores, which may be the reason patients 250 and 258 were identified as outliers.

Muscle Score

For muscle score, the percentage of patients with missing values was 21.7%, which is higher than for the heart and wheeze scores. The mean muscle score in both treatment groups was found to be lower than the other subcomponents. Figure 7.9 shows a plot of the mean dropout profiles in each group.

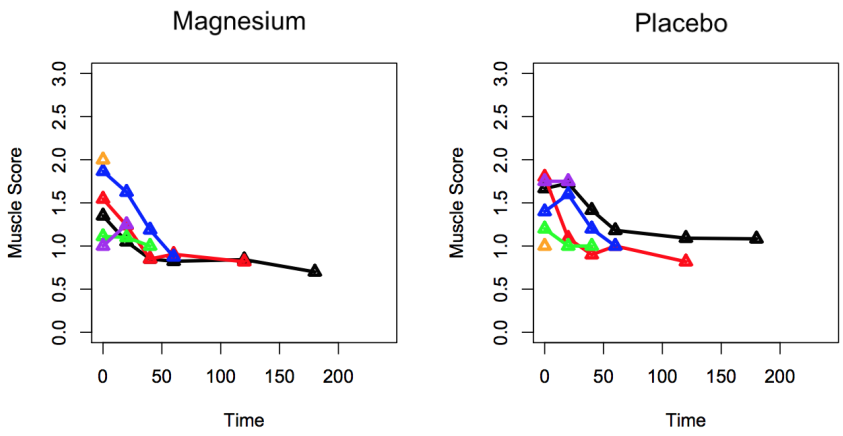


Figure 7.9: Mean Dropout Profiles for Muscle Score are each time point

For magnesium, all dropout groups had a decreasing muscle profile with the exception of those patients that dropped out at $t = 20$. For the placebo, all mean dropout profiles were decreasing prior to withdrawal - with the exception of dropouts at $t = 40$. There did not appear to be a substantial difference in the mean dropout profiles between treatments, however patients with a lower longitudinal profile appeared to be more likely to dropout.

Figure 7.10 shows a plot of the sequential discrimination to analyse the difference between dropouts and non dropouts.

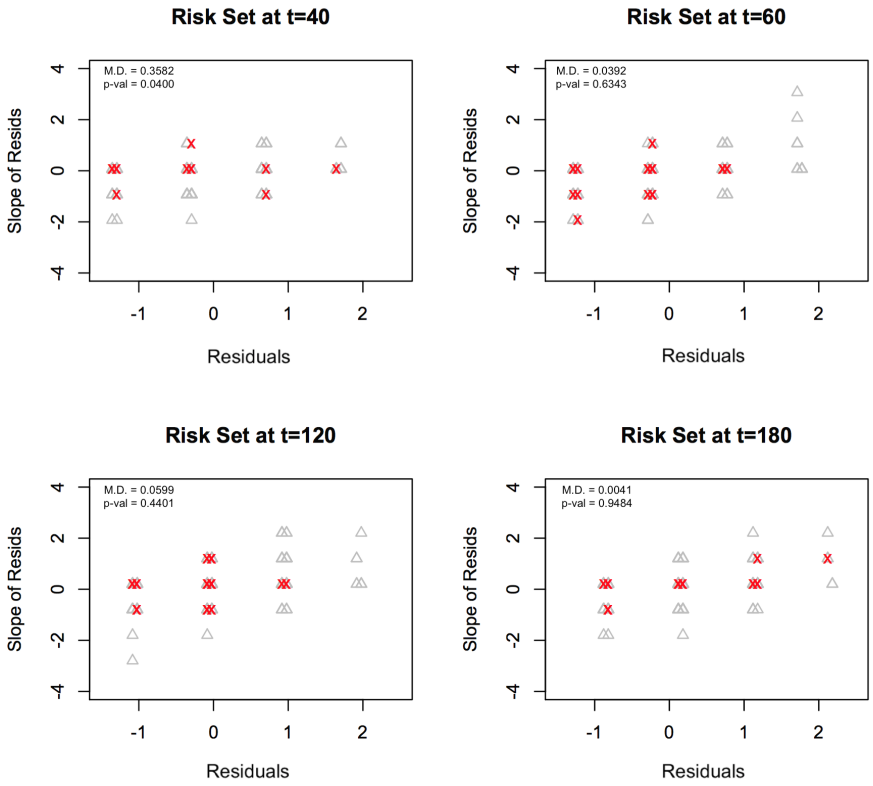


Figure 7.10: Sequential Discrimination for non-dropouts (grey triangles) v. dropouts (red crosses) at each time point

From the plots, a difference in residual profiles is observed at $t = 40$, with the majority of dropouts having a negative residuals or negative change in residuals, and the Mahalanobis distance was significant. No significant difference was found between dropouts and continuers at the other time points. However, due to the residual differences at $t = 40$, joint modelling methods may be appropriate. Table 7.3 details the results of the joint model fit.

Component	Parameter	Estimate	SE	95%Lower	95%Upper	P-Value
Longitudinal	(Intercept)	1.496	0.0421	1.418	1.575	
	time	-0.0034	1.8×10^{-4}	-0.0038	-0.0031	< 0.001
	β_1	-0.0764	0.0520	-0.169	0.0246	0.142
Survival	β_2	0.674	0.190	0.305	1.111	< 0.001
Association	γ	-0.369	0.192	-0.772	-0.0428	0.047
Variance	U_1	0.400	0.0308	0.335	0.449	
	U_2	8.0×10^{-6}	9.4×10^{-7}	6.2×10^{-6}	9.4×10^{-6}	

Table 7.3: Results for joint model fit to the MAGNETIC muscle score

Fitting the joint model did not show a significant longitudinal treatment effect (p=0.142), although on average patients randomised to magnesium had a lower muscle score (a parameter estimate of -0.0764). Dropout was significantly higher in the magnesium treatment group (p¡0.001), as was the case with both heart score and wheeze score. The significant γ estimate indicated that patients with lower muscle scores were more likely to drop out from the study. With 95% certainty the value of γ was found to be within [-0.772,-0.0428]. The following plot in Figure 7.11 shows the Cook's distances to identify influential cases.

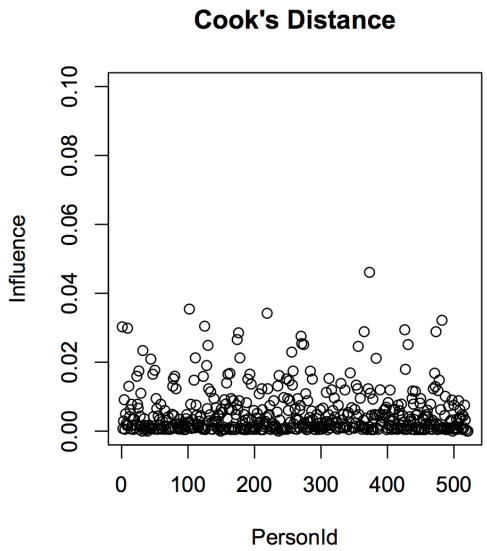


Figure 7.11: Plot of Influence for Each Patient

When observing the plot of Cook's distance, no obvious outliers were identified. The

most influential patient was number 373, who was also one of the most influential when the model fit in Section 2.5.2 was applied to ASS.

MAGNETIC Summary

In Chapter 2, an analysis of Asthma Severity Score for MAGNETIC yielded results which showed that patients randomised to magnesium generally had a lower severity score than in the placebo group. Patients randomised to Magnesium were also more likely to drop out than the placebo patients and individuals with lower severity scores were more likely to drop out. For the longitudinal outcomes of subcomponents, children in the magnesium group were found to have a significantly lower wheeze score than in the placebo group. However, the difference in outcome for individuals randomised to magnesium was not significant for the other components. In particular the parameter estimate for β_1 was only -0.01 for heart score. These results would suggest that adding magnesium to the standard nebuliser has a clinical impact on the ability of children's breathing abilities, but has very little effect on the heart rate.

When modelling wheeze score and heart score, the γ parameter was not found to be significant. However, γ was significant and negative for muscle use. As many children dropped out of the MAGNETIC trial due to a good prognostic outcome, the joint model results may imply that this is due to an observable improvement in muscle use abilities for certain patients. However, the magnitude of the γ parameter estimate was similar for each subcomponent, which does not support this hypothesis.

7.3 Mental Health Trial Data

In the MAGNETIC trial, a longitudinal outcome was modelled alongside time-to-dropout and all patients that completed the measurement schedule were treated as censored. How-

ever, in many trials the primary interest is not exclusively time-to-dropout, but time to a given clinical event or informative dropout. In the mental dataset as found in the joineR package in R [46], a longitudinal continuous outcome and time to informative dropout are recorded with censoring for patients that dropped out for non-informative reasons.

The data is taken from a trial involving mental health patients, in which 150 patients with chronic mental health issues were randomised to either placebo or one of two active interventions (denoted Treatment 2 and Treatment 3) with 50 patients in each treatment group. A continuous mental health score was recorded at baseline, as well as at 1,2,4,6 and 8 weeks post randomisation, and time to informative dropout was also recorded. In total, 63 (42.0%) of patients informatively dropped out from the study. To illustrate the applicability of these methods to balanced longitudinal data with an event outcome and intermediate censoring, the mental data set is analysed using methods similar to those described in Section 7.2.

Initially, a linear mixed model is fit to the longitudinal data and plots of sequential discrimination generated, as shown in Figure 7.12.

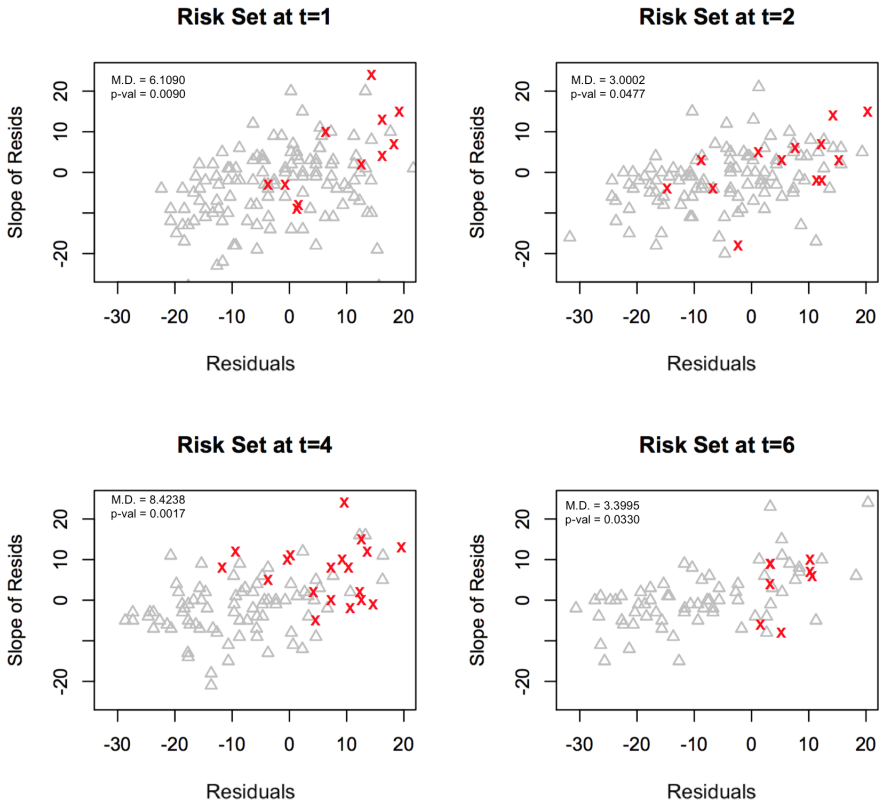


Figure 7.12: Sequential Discrimination for patients that remained in the trial (grey triangles) v. informative dropouts (red crosses) at each time point

Due to the continuous nature of the outcome, a difference in profiles between dropouts and non-dropouts is easier to observe than for the discrete data of the MAGNETIC analyses. From observing graphs, there appears to be a difference between dropouts and non-dropouts at most of the time points, with dropouts generally having both higher residuals on average and a higher residual slope than the patients who remained in the trial. An analysis of the Mahalanobis distance shows that there is a highly significant difference at $t = 1$ and $t = 4$, and a significant difference at $t = 2$ and $t = 6$. Therefore, joint models can be fit to the data. This initially incorporated a time variable, α, in the model. However this was not found to be significant, so another model was fit excluding this variable. Joint modelling was used to compare the two active treatments with placebo. $\beta_{1,1}$ refers to the

longitudinal comparison between placebo and treatment 2, while $\beta_{1,2}$ is the longitudinal comparison between placebo and treatment 3. Likewise for $\beta_{2,1}$ and $\beta_{2,2}$. Table 7.4 shows the results of the model fit.

Component	Parameter	Estimate	SE	95%Lower	95%Upper	P-Value
Longitudinal	(Intercept)	55.710	1.866	51.140	58.921	
	$\beta_{1,1}$	-1.322	2.571	-6.705	3.568	0.607
	$\beta_{1,2}$	-4.780	2.346	-9.309	-0.357	0.042
Survival	$\beta_{2,1}$	-0.522	0.402	-1.415	0.245	0.194
	$\beta_{2,2}$	-0.924	0.509	-2.156	-0.0771	0.070
Association	γ	0.0949	0.0186	0.06870	0.141	< 0.001
Variance	U_1	89.686	11.514	63.631	107.111	
	U_2	3.429	0.8846	1.846	5.658	

Table 7.4: Results of joint model fit to the mental trial

In the model, the two active treatments were compared to placebo. It was found that the longitudinal treatment effect for treatment 3 was a significant improvement on placebo (p=0.042); with the treatment effect being within [-9.303,-0.357]. However, treatment 2 was not found to be significant (p=0.607). There was not found to be a significant difference between treatment groups when modelling the hazard (p=0.194, 0.070), although patients in the placebo group dropped out informatively more frequently than in the active treatment groups, with parameter estimates of the log hazard ratio of -0.522 and -0.924 respectively. The estimate of γ was positive and significant (p¡0.001), demonstrating that patients with higher mental scores were more likely to experience a clinically informative drop out, which concurs with the plot in Figure 7.12. The confidence interval for γ showed that with 95% certainty the value of γ was within [0.069,0.141]. Figure 7.13 shows the influence of different patients in the trial.

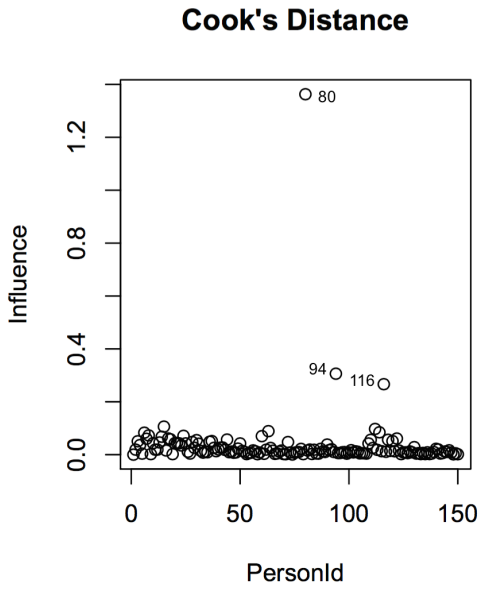


Figure 7.13: Plot of Influence for Each Patient

From the Cook's distance it can clearly by observed that patients 80, 94 and 116 have a higher influence than the other patients, and these are identified as outliers. The longitudinal profiles for the three most influential patients are shown in Figure 7.14.

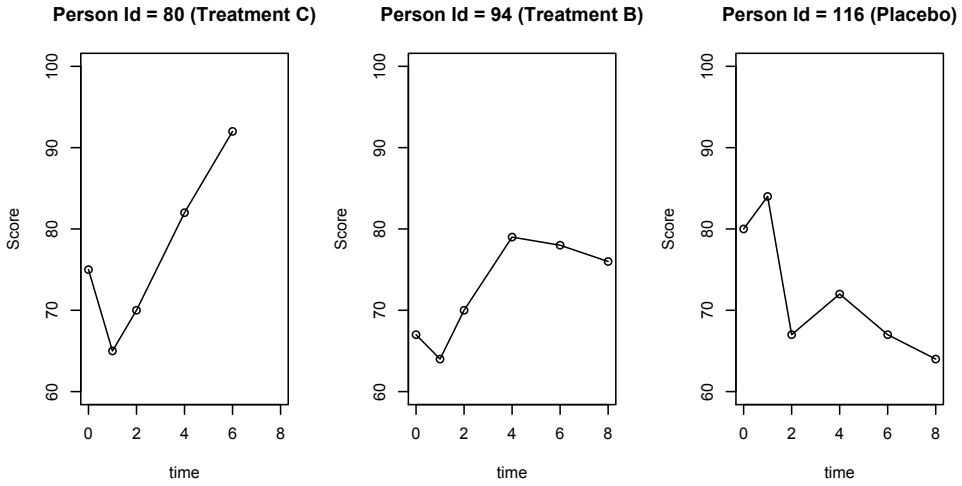


Figure 7.14: Longitudinal Profiles of Influential Cases

The influential nature of patient 80 with regard to the model is easy to observe from the graph of longitudinal outcome. While the majority of patients had a lower mental score in treatment group 3 than the other groups, patient 80 had a mental score increasing to 92, which was the highest outcome in the whole dataset. The outcome is consistently increasing after $t = 1$ up until the time of dropout. The influence of patients 94 and 116 may have been higher than other patients due to patient 94 having an increasing outcome from baseline despite being in an active treatment group, and patient 116 completing the measurement schedule and having a decreasing outcome despite being in the placebo group.

Analysis of Mental Health Trial Summary

In general, there was found to be no significant difference in mental score or hazard for patients in treatment group 2 when compared to placebo. However, patients in treatment group 3 had a significantly better outcome than the placebo group. This was an improvement of 4.78 points on average. Patients randomised to treatment 3 were also less likely to experience an informative dropout. The positive value of γ indicates that patients with higher mental scores were more likely to drop out informatively. This may seem intuitive as higher values imply a worse prognosis.

7.4 Liver Trial Data

To avoid the problem of missing data, some longitudinal trials have an unbalanced format [22]. In many trials, longitudinal readings are collected at different times for different patients [141] - [143]. This was not the case for MAGNETIC or the mental health trial, however joint modelling methods can be adapted and applied to this type of data. Unbalanced joint longitudinal and survival data is taken from a liver trial that was originally analysed in a non-joint modelling context [56].

In this trial, 488 patients with liver cirrhosis were randomised to either placebo or

prednisone and a measure of liver function, prothrombin index, was taken at baseline and then at different times for each patient. Prothrombin index is a discrete indirect marker of severe liver fibrosis [144] . The maximum time after randomisation that a patient's prothrombin index was measured in this trial was 11.1 years. Survival time was also measured and modelled as the time-to-event outcome.

Before the methods described in 7.2, 7.3 were applied to the liver trial, an exploratory data analysis was initially carried out. Prothrombin index scores ranged from 6 to 170 for each patient with a mean of 5.07 readings taken per subject, and the median follow up time for the study was 2.6 years. Of the 488 patients recruited to the study, there were 292 deaths (59.8%) and the median survival time was 2.1 years. An exploratory histogram of the survival times is plotted in Figure 7.15.

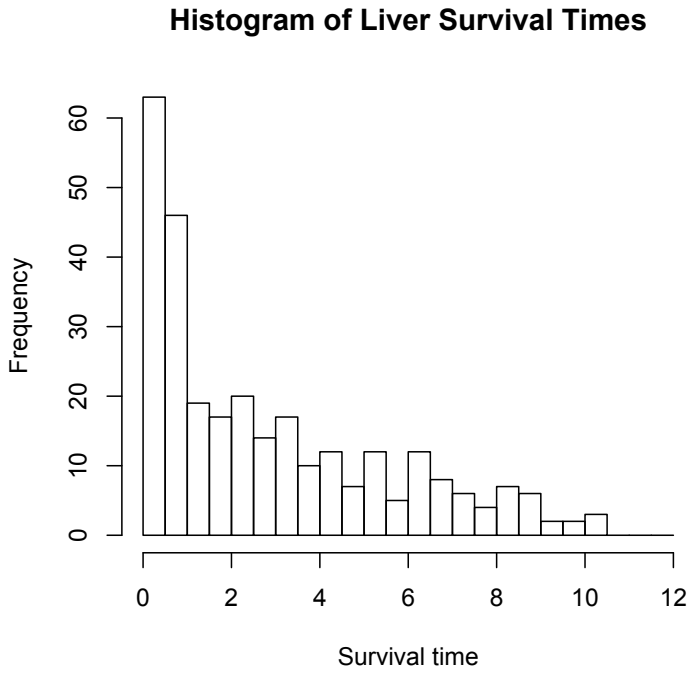


Figure 7.15: Plot of Censored Survival Times

In the trial, 37.0% of patients that experienced an event did so in the first year, and 21.6% within the first 6 months. The median survival in the placebo group was 1.773 years, while this was longer in the prednisone group (2.319 years). Figure 7.16 shows the LOESS smoothed function of prothrombin index for patients randomised to placebo and prednisone prior to death.

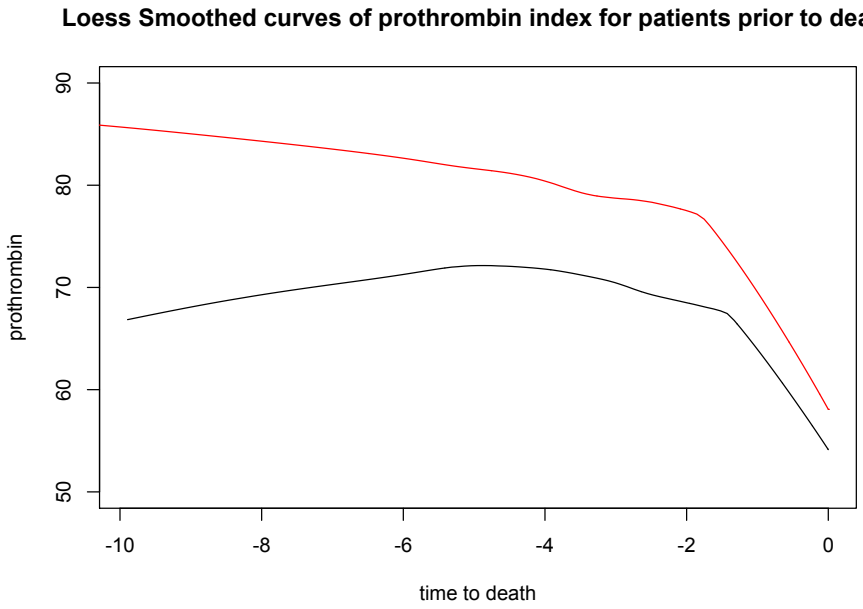


Figure 7.16: Placebo = Black line, Prednisone = Red Line

The mean change in prothrombin index was decreasing in both treatment groups prior to dropout, which appears to verify that the outcome is a successful biomarker. Rapid decreases in prothrombin index appears to occur at around 2 years prior to death in the prednisone group and 20 months in the placebo group. Also the index was found to be higher in general in the prednisone group for patients that died than the placebo group. Analysis of the outcome for censored patients showed that the mean prothrombin score was 88.369, which can be observed from Figure 7.16 as higher value than for the patients that died.

In Figures 7.2, 7.6, 7.10 and 7.12, sequential discrimination was used for each timepoint to observe the difference in residual profiles for patients that experienced an event and the ones that did not experience an event. In an unbalanced longitudinal trial, the time points are not pre-specified or the same for each patient. Therefore we propose an amendment to this method for unbalanced trials. To carry out the sequential discrimination, initially fit a linear mixed model to the data as described in Chapter 6.2.1. For all patients who had at least two longitudinal readings taken, calculate the residual for their last longitudinal reading prior to death/being censored and also the change in residual from the previous measurement. These can then be plotted and the patients that experienced death identified graphically. Mahalanobis distance will again be used to quantify the difference in residual profiles for the censored and non-censored patients. Figure 7.17 shows this plot.

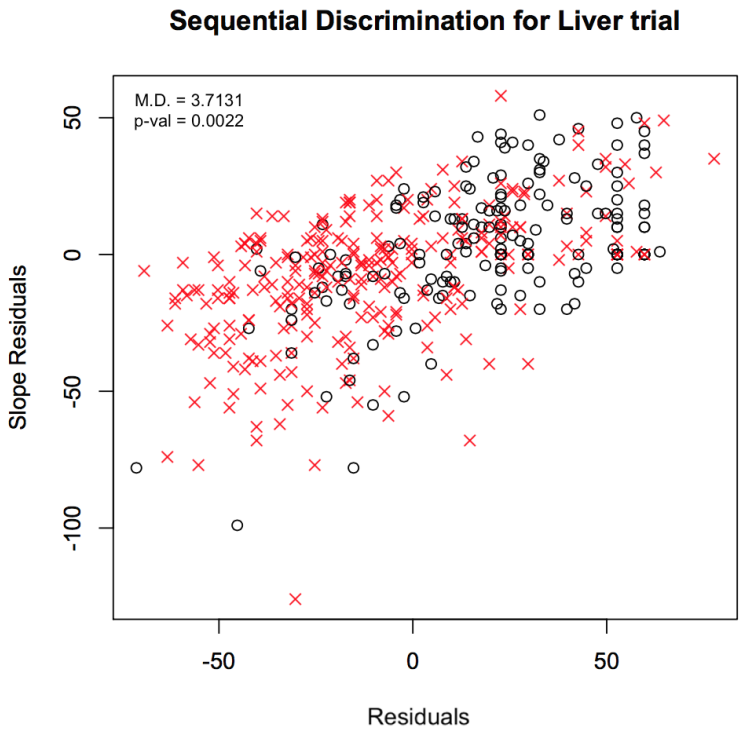


Figure 7.17: Sequential Discrimination Plot; Censored = Black (o), Died = Red (x)

The residual profiles and slope of residuals for patients that died seems to be lower on average than for the censored observations, with the majority of censored observations having positive values for both elements. An analysis of Mahalanobis distance showed that the difference between the two groups was highly significant (a p-value of 0.0022), therefore implying joint modelling may be appropriate.

Initially a random slope and intercept joint model was fit with time and treatment variables in the longitudinal component, and treatment in the survival component. However, the time variable was not found to be significant. Therefore another joint model was fit without α, and the results are given in Table 7.5.

Component	Parameter	Estimate	SE	95%Lower	95%Upper	P-Value
Longitudinal	(Intercept)	69.990	1.264	67.274	72.623	
	β_1	6.929	1.817	2.916	10.431	< 0.001
Survival	β_2	-0.0962	0.140	-0.363	0.173	0.492
Association	γ	-0.0416	0.00397	-0.0504	-0.0345	< 0.001
Variance	U_1	354.498	26.57	303.001	406.220	
	U_2	16.809	4.052	11.010	25.372	

Table 7.5: Results of joint model fit to liver data

Table 7.5 shows that patients randomised to prednisone had on average a prothrombin index of 6.929 higher than in the placebo group, which was found to be a significant difference(p¡0.001). The treatment effect was estimated to be within the 95% confidence interval of $[2.616, 10.431]$. The log hazard ratio between the treatment groups was not found to be significant (p=0.492), although on average patients randomised to the placebo group were more likely to experience death in this study (a log hazard ratio of -0.0962). Patients with a lower prothrombin index were also more likely to experience death, as γ was found to be significant and negative (p¡0.001). The parameter of γ was estimated to be within $[-0.050, -0.035]$ with 95% certainty.

Methods of Cook's distance as described in 6.2.3 can be applied to this model despite the unbalanced format. Figure 7.18 shows the results of these diagnostics to identify outliers.

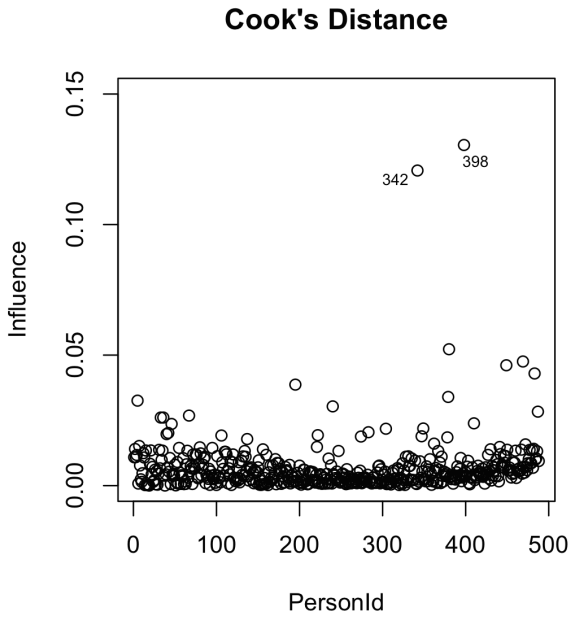


Figure 7.18: Plot of Influence for Each Patient

From calculating the influence it was determined that patients 342 and 398 were outliers, as they had Cook's distances greater than 3 times the mean of the Cook's distance for all patients. Figure 7.19 shows the longitudinal profiles for these patients.

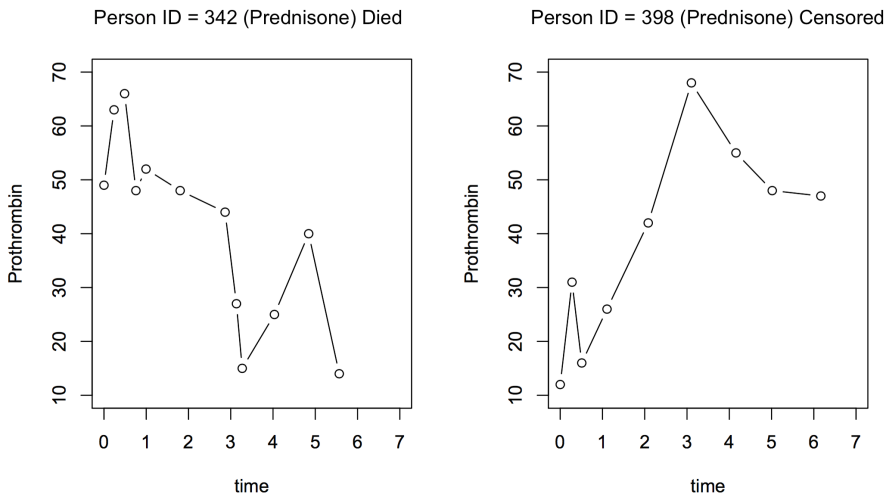


Figure 7.19: Longitudinal Profiles of Influential Cases

Both outliers were randomised to prednisone. Patient 342 died in the 6th year after randomisation, which is approximately 4 years over the median for a subject prescribed to prednisone. Also, observing Figure 7.17 we can see that this patient had a prothrombin index profile lower than the mean for patients in prednisone. The patient's liver function fluctuated greatly between 2 and 6 years with a swift decline in prognosis detected at the start of year 3 before a temporary recovery.

At baseline patient 398 had a very low prothrombin index which remained below 35 for the first 2 years. However, after this time the patient experienced an improvement in liver function, with the prothrombin index reaching a peak of 69 before the patient was censored in the 7th year post randomisation. It may be expected that a patient with such a low prothrombin index at baseline would have a low probability of survival, particularly as on average the prothrombin index for all patients was decreasing over time.

Liver Data Discussion

By fitting the random slope and intercept joint model to the liver data set, conclusions can be easily drawn from the results of the model fit which may have required a multitude of more time consuming analyses to establish. In general, prednisone increased the prothrombin index for patients and therefore their liver function. Patients randomised to prednisone had a smaller hazard of death, but this was found not to be significant. The analysis revealed that patients with a lower prothrombin index were at greater risk of death, which can be verified by Figure 7.20. This confirms the success of prothrombin index for liver cirrhosis patients as a biomarker for death.

7.5 Discussion

In this chapter, the methods discussed throughout this thesis have been applied to a more diverse range of data sets in order to display their versatility. The hope is that the reader will gain an understanding of how these methods can be used to draw conclusions in RCT analyses. It was demonstrated that sequential discrimination methods are an effective tool for establishing a need for joint modelling. These methods were particularly useful for continuous outcome data as graphical methods were easier to interpret. In section 7.4, a method for using sequential discrimination in an unbalanced trial design was proposed. This method established a significant difference between patients that died and the censored observations, suggesting joint modelling methods would be appropriate. This turned out to be the case as γ was found to be highly significant once a joint model had been fit.

Now that joint modelling has been presented in detail, a summary of these investigations, drawbacks and ideas for future work are presented in Chapter 8.

Chapter 8

Discussion

8.1 Introduction

The focus of this thesis has been on the development of research in the area of joint modelling of longitudinal and time-to-event data, and the aim of the research was to develop a greater understanding of how these methods can be applied. The joint model as specified by Henderson [10] uses latent Gaussian variables to monitor a longitudinal outcome alongside the time to an event. In many clinical trials with longitudinal outcomes, there will inevitably be patients that dropout or fail to complete the measurement schedule, and in these cases analysing data by utilising joint models is an efficient method. However, the systematic review conducted in Chapter 3 showed that joint modelling of longitudinal and time-to-event data is rarely used in practice, which may be a result of a lack of understanding or awareness of these techniques when conducting a trial and analysing trial data. The research conducted within this thesis should provide a greater understanding of the joint modelling framework to the reader, and aid with the development and application of joint modelling when conducting an RCT. This thesis provides practical examples of when joint modelling is utilisable in a trial, derived power and sample size formulae for studies involving joint models, an understanding of the association between longitudinal and event time outcome and development of software to implement diagnostic methods for joint models.

8.2 Topics Covered

In Chapter 1, an overview of conducting an RCT was presented and the importance of good trial design was highlighted. Methods of randomisation, blinding, sample size calculation and the measurement schedule all should be discussed in the planning stage of a trial, both for RCTs where standard analyses are used and for trials that employ joint modelling. One major issue to be addressed in the trial design is how to handle missing data, as a mishandling of the outcomes for patients that drop out from a trial can result

in misleading results. The joint modelling framework was also introduced in this chapter as a method for missing data handling, and the MAGNETIC trial was introduced, which was the main motivation behind this thesis.

A description of some commonly used methods for handling missing data were discussed in Chapter 2, as well as the potential advantages and disadvantages of employing such methods in a longitudinal data context. A unique simulation study was carried out to compare the performance of existing missing data handling and imputation methods to joint modelling for estimating the treatment effect (β_1). When the missingness is not MCAR, it was concluded that joint modelling performed consistently well for data with $20, 30, 40$ and 50% dropout. Joint modelling was more successful for estimating β_1 than complete case and LOCF based methods, and showed a similar level of accuracy to the multiple imputation methods used. Joint modelling results also showed a relative bias for β_1 of under 2.1% in all cases.

A practical application of the random slope and intercept joint model was also demonstrated by using the MAGNETIC data, in which we aimed to test the benefit of adding magnesium nebuliser solution to the standard treatment for children with acute sever asthma. When applying joint models, this analysis indicated that adding magnesium nebuliser solution to the standard severe asthma treatment resulted in a significantly lower ASS score, and that children randomised to magnesium dropped out more than in the placebo group. This difference was not detectable when using a complete case analysis alone, which highlighted the benefit of joint models in this scenario. The association parameter, γ, was estimated to be -0.18. This indicates that patients with lower ASS profiles were more likely to leave the study, which may have been due to these children feeling better, and hence being ready to be discharged from the hospital.

A systematic review was carried out in Chapter 3 to establish what methods were

used for missing data handling in trials with longitudinal outcome data between 2005 and 2012 [60]. This study showed that many papers fail to acknowledge the issue of missing data handling, and that the most commonly used method of missing data handling was a complete case analysis. Of the trials with missing data, 31.8% made no references to the reasons for dropout in the article. Imputation methods were used in 18.0% of trials. On a positive note, the results indicate that missing data problems have been addressed more in recent years. To encourage transparency in trial reporting of missing data, a Four Point Plan was proposed to trialists as guidelines for missing data handling. This plan contained missing data reporting standards that should be adhered to, encouraged the inclusion of a discussion of the pros and cons of the missing data methods used in each trial and suggested that all results for imputed data should be presented alongside a complete case analysis. This study also highlighted that few RCTs employ joint modelling, as no studies in this systematic review used these methods.

In Chapter 4, an investigation into the sample size formulae for the joint model parameters β_1, β_2 and γ was conducted, and sample size formulae for γ and β_2 were derived for the random slope and intercept joint model. For the joint model specification by Henderson [10], no published work has currently focused on sample size and power calculations. It is suggested that this may be one of the reasons that joint modelling is rarely used for primary analyses in clinical trial literature. For β_1, simulations demonstrated that increasing the number of time points and lower values of $Var(U_1)$ results in higher powers. The effect of varying γ and $Var(U_2)$ on power was found to be negligible for β_1, although it was found that the power when using joint models was significantly higher than for a complete case analysis. A sample size formula was derived for β_2 in the case where $\gamma = 0$ using the Rao score statistic. This formula was then adjusted for $\gamma \neq 0$, as it was discovered that γ and $Var(U_2)$ have an impact on the number of required events. Specifically the simulation study showed that as γ increases, the power for β_2 decreases and the magnitude of this change is proportional to $\gamma^2 Var(U_2)$. Similarly, a sample size formula for γ was

derived using the Rao score statistic. The power for γ is dependent on the truncated moment for the distribution of dropout, $Var(U_1)$, $Var(U_2)$ $Cov(U_1, U_2)$. An approximation to the truncated moment required to estimate the sample size is proposed based upon the uniform distribution, which takes into account the number of longitudinal time points when calculating power. The simulation study showed that the this approximation performed well, and that the number of time points in a study had an impact on the power for γ. Although the derived sample size formulae rely on knowledge of variances of the random effects, for many phase III trials with time-to-event data, preliminary studies are undertaken to assess potential side effects and success of the treatments. By applying joint models to the preliminary data, it is possible to approximate values for $Var(U_1)$,$Var(U_2)$ and $Cov(U_1, U_2)$ prior to a larger scale trial. Using these estimates, sample sizes for the γ and β_2 parameters for joint models can be approximated using the formulae derived in Chapter 4.

Despite the discussions and reliance upon the γ parameter in joint models, an investigation of the properties of γ has not been conducted in published literature, beyond the definition for a negative and positive value. Properties of different magnitudes of γ and a visualisation for different γ values is provided in Chapter 5 using simulated data. This work showed that γ is dependent on the change in longitudinal outcome pre-dropout, the percentage dropout (baseline hazard) and $Var(U_2)$. This is demonstrated through the development of an approximation formula to model relationship between γ and change in outcome pre-dropout for a commonly used trial design.

Another area in which little work has been done in joint modelling is diagnostic procedures. Only two papers have proposed diagnostic methods for joint modelling [119, 120], and these methods have not been used in practice to analyse trial data. In Chapter 6, software was developed in R to apply appropriate diagnostic methods to the random slope and intercept joint model, and the self-made functions for sequential discrimination and Cook's

distance will be included in the next update of joineR. A new variation of sequential discrimination which focuses solely on the longitudinal profiles immediately prior to dropout was proposed for testing the appropriateness of joint modelling by observing differences in longitudinal profiles between patients that left the study and those that completed the measurement schedule at each time point. Using this new method, a pre-analysis sequential discrimination for the MAGNETIC trial data indicated that differences between these two groups of patients existed, and therefore joint modelling methods were appropriate in this instance. Cook's distance in joint modelling was used to identify the subjects with the greatest influence. For joint models with a parametric time-to-event specification, a residual analysis based upon multiple imputation can be used to assess model fit. A demonstration of this is also provided in the chapter.

In Chapter 7, the applicability of joint modelling to many different trial scenarios was illustrated. This work provides a guide for using and interpreting joint models to trialists and statisticians. However, for someone who is more interested in the methodological elements of joint modelling, the derivations in Chapter 4 and formulation in Chapter 5 provides greater context to the work.

The work presented in this thesis emphasised that joint modelling can be a useful tool for modelling both a longitudinal outcome and dropout simultaneously. In many trials there are clinical benefits for assessing and comparing the time-to-dropout between treatment groups, and also estimating the relationship between longitudinal outcome and dropout. To obtain this information, without using joint models, would be a time consuming procedure, requiring graphical representations to coincide with multiple model fits. With literature available on model formulation, the work in this thesis on γ parameter visualisation, sample size calculations and software development will contribute to the understanding of how to use this type of modelling, and therefore encourage trialists to apply joint models when appropriate.

8.3 Limitations

Within this work, the MAGNETIC trial has acted as a motivating dataset and in Chapter 7 the methods described were applied to a wider range of trial scenarios. However, this is still only a limited number of datasets to test the proposed methods. Unfortunately the nature of missing data in a longitudinal trial will always have elements of uncertainty, and some patients will inevitably have intermittent missing values. The simulation studies in this thesis have not taken into account the effect of missing intermittent data as this would complicate the analyses and interpretations.

In Chapter 3, a systematic review was carried out and data extracted from 100 papers selected at random due to time constraints. While this provides a snapshot of the frequency that each missing data method is used, to obtain the most accurate data all papers highlighted by the search could have data extracted and analysed. Therefore while no papers within the sample were found to have used joint modelling of longitudinal and time-to-event data, it is possible that the ones that did were not part of the random selection. For this work to be done in the future, a full list of the 381 papers is available on request, although it is advised than an updated systematic review is carried out with papers included from 2012 onwards.

In Chapter 4, sample size formulae were generated for the γ and β_2 parameters in joint modelling based upon knowledge or approximations of γ, $Var(U_1)$ and $Var(U_2)$, which may be difficult to estimate prior to a trial commencing. In this case, Bayesian methods can be used to calculate sample size and $Var(U_1)$, $Var(U_2)$ which can prove to be a difficult and time consuming task for trialists [40, 48]. Also for β_1, due to complexity, the sample size and power calculations must be carried out using simulations. The majority of the work in this chapter is based upon approximations and is calculated using datasets with equally spaced time points. In practice, time points are not also equally spaced which

may decrease the accuracy of the approximate sample size formulae.

As discussed in Chapter 5 Section 1, when visualising the relationship between γ and change in outcome prior to dropout, knowledge of $Var(U_2)$ is required. While in many trials an approximation to this value is available from preliminary studies, this may not always the case. Also due to the nature of the gaussian variables within model means there are limitations on certain γ values for a given $Var(U_2)$, and higher magnitudes of γ can be difficult to simulate.

Finally, for the newly developed variation of sequential discrimination, it can be difficult to see the differences in profiles from the plots alone when a longitudinal outcome is discrete, as can be observed in Figure 6.1. In order to overcome this, a 3 dimensional graphical representation of the results could possibly be used to indicate the number of data points on each discrete value. This is something to consider for the future, as a standard 3 dimensional graph may appear convoluted and difficult to discern. However, this discrete data does not reduce the effectiveness of using Mahalanobis' distances to generate a p-value.

8.4 Future Work

While this thesis has contributed to the understanding of the joint modelling framework proposed by Henderson et al. [10] as well as the trial design considerations, there is still a wide range of possibilities for further research. As highlighted in Section 8.3, the simulated datasets fail to include intermittent missing data or unbalanced datasets. The effect of intermittent data caused by patients missing visits is a common occurrence that has not yet been addressed in literature for joint modelling. To investigate the effect of adjusting for intermediate missing data in a trial we can begin by analysing the MAGNETIC dataset, in which intermediate missingness was present, using three separate methods; the standard

analysis which treats dropout as time to first missing value as described in Chapter 2.4 of this thesis, an analysis which imputes values for the intermittent missing data before fitting joint models and a third analysis which treats MAGNETIC as an unbalanced study design with each individuals longitudinal measurements being taken at different time points for different patients. We can compare the results and establish whether the conclusions are the same for each analysis. To investigate this further, a simulation study can be used in which different patterns of intermittent missingness are used for simulated datasets from a joint model. The aforementioned analyses are then employed and the differences in parameter estimates and confidence intervals are assessed. This way, we can also assess different methods of multiple imputation for imputing the intermittent values.

In terms of the sample size formula and power for β_1, an accurate approximation is yet to be derived in the literature despite many trials having the analyses of an outcome observed over time as the primary goal in a study. Currently simulations are the most accurate method of estimation. One of the main difficulties with deriving such a formula and calculating the power is that the patterns of missing data will have an effect on the power, and the research on this topic has not been covered extensively. The missingness patterns and properties within a trial can be assessed during the analysis, and these patterns will not be known in the design stage. However, in the case where time to dropout is analysed, this becomes less of a problem as the number of missing values will follow a less complicated pattern. The simulation study in Section 4.5 illustrated that γ has no impact on the power of β_1. Therefore patterns of dropout would be the focus of this work. A sample size formula for a linear mixed model is proposed in Hedeker et al (1999) [110] which is also appropriate for the β_1 parameter in Joint Models. However this calculates the total number of patients required to be remaining in the study at each time point. This may not seem appropriate for a trialist, as predicting the number of dropouts or events at each time point prior to a trial commencing requires some guesswork. To address this problem, I would propose a more extensive simulation study. Building on the simulations

carried out in Chapter 4.5, by using a greater variation of trial properties, the aim would be to establish a simpler sample size approximation formula that can be used by trialists in the design stage of a trial. In such a study the median length of follow up, the distribution of dropout times, percentage of missingness across treatment groups, number of time points and $Var(U_1)$, $Var(U_2)$ could be varied. With these large numbers of different trial properties, it may be possible to accurately approximate the sample size in a longitudinal trial with time-to-dropout included for β_1.

Additionally, further simulations can be carried out to assess the best method of approximating the VIF for the power of β_2. In Chapter 4 an approximation for the VIF was calculated for the MAGNETIC trial design, but this approximation formula may not be valid for all trial designs. A further simulation study based on the work done in this Chapter can be used to test this relationship for different RCT scenarios. Currently, no work of this nature has been carried out in published literature, and no sample size formula which includes a patient specific random slope to model the hazard has been derived.

In the MAGNETIC dataset, the ASS score was recorded for children over the period of 240 minutes at 7 different time points. At each time point the physician rated the children's ability to breathe, their heart rate and their muscle function out of 3, and the combined scores make up the ASS. In Section 7.2, these three components were analysed individually and it was discovered that the wheeze score was significantly different in the treatment groups. However, γ was found to be significant only in the analysis of the heart rate data. In practice, we are aware of how these three separate variables compose together to make the ASS score. However, we are unaware of how the wheeze score correlates with heart score/muscle score without performing separate correlation analyses which may fail to take into account certain elements of the joint model. Therefore it would be of great benefit to develop a joint modelling framework capable of monitoring multiple outcomes over time in a multivariate setting for the Henderson specification of the joint model. Fiews

and Verbeke (2004) [145] discussed some of the pitfalls of using random effects to estimate multivariate longitudinal profiles and this paper can provide a guide and motivations to establish a multivariate joint modelling based upon the Henderson specification of the joint model [10]. Some work has been done in this area [52, 146], but the majority of the literature focuses on Bayesian methods [49, 51]. Frequentist multivariate joint modelling is yet to be investigated to the point where most trialists feel comfortable using the methods.

More recently, Verbeke et. al (2014) [147] presents a random effects model for longitudinal multivariate data which takes into account the marginal cross correlations between mean zero random effects within each longitudinal outcome. The aim for the multivariate joint model development would be to establish a way to jointly use this model alongside the Henderson specification of the joint model, for which all outcomes will have the same event times. This way, multiple γ's could be estimated and details of the correlations between longitudinal joint model estimates investigated. While this model is yet to be developed, I would propose the use of 3-dimensional latent mean zero Gaussian variables as a basis to connect the separate components of this multivariate joint model.

8.5 Concluding Remarks

Joint modelling of longitudinal and time-to-event data is an under-utilised methodology that allows conclusions to be drawn about longitudinal treatment effect, dropout and details of patient prognosis prior to dropout. It has been shown that in appropriate situations, this methodology can prove to be an efficient and accurate statistical tool. However, research in this area is currently in the developmental stage. The work presented in this thesis has provided a contribution to the understanding of this topic, and will encourage researchers to explore joint modelling further.

Bibliography

[1] Moher, David et al. "CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMJ: British Medical Journal.* 340 (2010): c869.

[2] Little, Roderick J., et al. "The prevention and treatment of missing data in clinical trials." *New England Journal of Medicine.* 367.14 (2012): 1355-1360.

[3] Gabriel, Andre P., and Charles P. Mercado. "Data retention after a patient withdraws consent in clinical trials." *Open Access Journal of Clinical Trials* 3 (2011): 15.

[4] Carpenter, James R., and Michael G. Kenward. "Missing data in clinical trials: a practical guide." *National Institute for Health Research*, Publication RM03/JH17/MK: Birmingham (2008).

[5] Fayers, Peter M., Desmond Curran, and David Machin. "Incomplete quality of life data in randomized trials: missing items." *Statistics In Medicine* 17.5:7 (1998): 679-696.

[6] Rubin, Donald B. "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91.434 (1996): 473-489.

[7] Schafer, Joseph L. "Multiple imputation: a primer." *Statistical Methods In Medical Research* 8.1 (1999): 3-15.

[8] Tseng, Yi-Kuan, Fushing Hsieh, and Jane-Ling Wang. "Joint modelling of accelerated failure time and longitudinal data." *Biometrika* 92.3 (2005): 587-603.

[9] Williamson, P. R., et al. "Joint modelling of longitudinal and competing risks data." *Statistics In Medicine* 27.30 (2008): 6426-6438.

[10] Henderson, Robin, Peter Diggle, and Angela Dobson. "Joint modelling of longitudinal measurements and event time data." *Biostatistics* 1.4 (2000): 465-480.

[11] Stanley K. "Evaluation of randomized controlled trials." *Circulation* 115.13 (2007):1819-22.

[12] Chalmers, Thomas C., et al. "A method for assessing the quality of a randomized control trial." *Controlled Clinical Trials* 2.1 (1981): 31-49.

[13] Collins, Joseph F. "Clinical Trials Protocols." *Encyclopedia of Biostatistics* (2005).

[14] Sackett, David L., et al. "How to practice and teach EBM." *Edinburgh: Churchill Livingstone* (2000).

[15] Matthews, John NS. "Introduction to randomized controlled clinical trials." *CRC Press* (2006).

[16] Harper, Diane M., et al. "Sustained efficacy up to 4-5 years of a bivalent L1 virus-like particle vaccine against human papillomavirus types 16 and 18: follow-up from a randomised control trial." *The Lancet* 367.9518 (2006): 1247-1255.

[17] Kendall, J. "Designing a research project: randomised controlled trials and their principles." *Emergency Medicine Journal: EMJ* 20.2 (2003): 164.

[18] Kayne, G. Gregory. "The use of sanocrysin in the treatment of pulmonary tuberculosis:(Section of Medicine)." *Proceedings of the Royal Society of Medicine* 28.11 (1935): 1463.

[19] Friedman, Lawrence M., Curt Furberg, and David L. DeMets. "Fundamentals of clinical trials. Vol. 4." *New York: Springer* (2010).

[20] NHS, "Clinical Trials and Medical Research - Phases of trials", http://www.nhs.uk/Conditions/Clinical-trials/Pages/Phasesoftrials.aspx, 14th September 2014.

[21] Smeaton, Laura M., "Introduction to Clinical Trial Design - Presentation", http://dsg.harvard.edu/courses/hst951/Spring02/951-23.pdf, 11th March 2015

[22] Diggle, Peter, Kung-Yee Liang, and Scott L. Zeger. "Longitudinal data analysis." (1994).

[23] Jackson, Dan, Ian R. White, and Morven Leese. "How much can we learn about missing data?: an exploration of a clinical trial in psychiatry." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.3 (2010): 593-612.

[24] Laird, Nan M. "Missing data in longitudinal studies." *Statistics In Medicine* 7.1:2 (1988): 305-315.

[25] Blehar, Mary C., et al. "Enrolling pregnant women: issues in clinical research." *Women's Health Issues* 23.1 (2013): 39-45.

[26] Drane, James F. "Competency to give an informed consent: a model for making clinical assessments." *Jama* 252.7 (1984): 925-927.

[27] Lachin, John M. "Introduction to sample size determination and power analysis for clinical trials." *Controlled Clinical Trials* 2.2 (1981): 93-113.

[28] Jones, B., et al. "Trials to assess equivalence: the importance of rigorous methods." *BMJ : British Medical Journal* 313.7048 (1996): 36-39.

[29] Christensen, Erik. "Methodology of superiority vs. equivalence trials and non-inferiority trials." *Journal of Hepatology* 46.5 (2007): 947-954.

[30] Suresh, K. P. "An overview of randomization techniques: an unbiased assessment of outcome in clinical research." *Journal of Human Reproductive Sciences* 4.1 (2011): 8.

[31] Torgerson D., "Allocation Methods - Presentation", www-users.york.ac.uk/ djt6/NIcourse/randomisation.ppt, 14th March 2015.

[32] Hahn, Seokyung. "Understanding noninferiority trials." *Korean journal of Pediatrics* 55.11 (2012): 403-407.

[33] McCulloch, Charles E., and John M. Neuhaus. Generalized linear mixed models. *John Wiley & Sons*, Ltd, 2001.

[34] Higgins, Julian PT, ed. "Cochrane handbook for systematic reviews of interventions." Vol. 5. Chichester, England: Wiley-Blackwell, 2008.

[35] Chan, An-Wen, and Douglas G Altman. "Identifying Outcome Reporting Bias in Randomised Trials on PubMed: Review of Publications and Survey of Authors." *BMJ: British Medical Journal* 330.7494 (2005): 753.

[36] British guideline on the management of asthma; *Scottish Intercollegiate Guidelines Network - SIGN* (Oct 2014)

[37] Powell CVE, Kolamunnage-Dona R, Lowe J, et al. ; on behalf of the MAGNETIC study group "MAGNEsium Trial In Children (MAGNETIC): a randomised, placebo-controlled trial and economic evaluation of nebulised magnesium sulphate in acute severe asthma in children." *Southampton (UK): NIHR Journals Library*; 2013 Oct. (Health Technology Assessment, No. 17.45.)

[38] Yung, M., M. South, and T. Byrt. "Evaluation of an asthma severity score." *Journal Of Paediatrics And Child Health* 32.3 (1996): 261-264.

[39] Rizopoulos, Dimitris. "Joint models for longitudinal and time-to-event data: With applications in R." *CRC Press*, 2012.

[40] Tsiatis, A. A., Victor Degruttola, and M. S. Wulfsohn. "Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4

counts in patients with AIDS." *Journal of the American Statistical Association* 90.429 (1995): 27-37.

[41] Lin, D. Y., M. A. Fischl, and D. A. Schoenfeld. "Evaluating the role of CD4:lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials." *Statistics In Medicine* 12.9 (1993): 835-842.

[42] Wulfsohn, Michael S., and Anastasios A. Tsiatis. "A joint model for survival and longitudinal data measured with error." *Biometrics* (1997): 330-339.

[43] Tsiatis, Anastasios A., and Marie Davidian. "Joint modeling of longitudinal and time-to-event data: an overview." *Statistica Sinica 14.3 (2004)* : 809-834.

[44] Faucett, Cheryl L., and Duncan C. Thomas. "Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach." *Statistics In Medicine* 15.15 (1996): 1663-1685.

[45] Sousa, Ines. "A review on joint modelling of longitudinal measurements and time-to-event." *Revstat* 9.1 (2011): 57-81.

[46] Philipson, Pete, Ines Sousa, and Peter J. Diggle . "joineR: Joint modelling of repeated measurements and time-to-event data.", http://cran.r-project.org/web/packages/joineR/index.html. (2012).

[47] Powell, C, Kolamunnage-Dona R, Lowe J, Boland A, Petrou S, Doull I, Hood K, Williamson P. "Magnesium sulphate in acute severe asthma in children (MAGNETIC): a randomised, placebo-controlled trial." *The Lancet Respiratory Medicine* 2013, 1 (4): 301-308.

[48] Chen, Liddy M., Joseph G. Ibrahim, and Haitao Chu. "Sample size and power determination in joint modeling of longitudinal and survival data." *Statistics In Medicine* 30.18 (2011): 2295-2309.

[49] Ibrahim, Joseph G., Ming-Hui Chen, and Debajyoti Sinha. "Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials." *Statistica Sinica* 14.3 (2004): 863-884.

[50] Chi, Yueh Yun, and Joseph G. Ibrahim. "Joint models for multivariate longitudinal and multivariate survival data." *Biometrics* 62.2 (2006): 432-445.

[51] Xu, Jane, and Scott L. Zeger. "Joint analysis of longitudinal data comprising repeated measures and times to events." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50.3 (2001): 375-387.

[52] Song, Xiao, Marie Davidian, and Anastasios A. Tsiatis. "A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data." *Biometrics* 58.4 (2002): 742-753.

[53] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal Of The Royal Statistical Society.* Series B (methodological) (1977): 1-38.

[54] Little, Roderick JA. "Modeling the drop-out mechanism in repeated-measures studies." *Journal of the American Statistical Association* 90.431 (1995): 1112-1121.

[55] Shizgal, B. (1981). "A Gaussian quadrature procedure for use in the solution of the Boltzmann equation and related problems". *J. Comp. Phys.* 41: 309?328.

[56] Andersen, P. K., O. Borgan, and R. D. Gill. N. Keiding. "Statistical Models Based on Counting Processes." *Springer Science & Business Media* (2012)

[57] Breslow, Norman. "Odds ratio estimators when the data are sparse." *Biometrika* 68.1 (1981): 73-84.

[58] Wothke, Werner. "Longitudinal and multigroup modeling with missing data." (2000).

[59] Little, Roderick JA, and Donald B. Rubin. "Statistical analysis with missing data." *John Wiley & Sons*, 2014.

[60] Powney, M., Williamson, P., Kirkham, J., & Kolamunnage-Dona, R. "A review of the handling of missing longitudinal outcome data in clinical trials". *Trials*, 15.1 (2014): 1-19.

[61] Rubin, Donald B. "Inference and missing data." *Biometrika* 63.3 (1976): 581-592.

[62] Pigott, Therese D. "Methods for handling missing data in research synthesis." *The Handbook of Research Synthesis* 4.10 (1994).

[63] Dziura, James D., et al. "Strategies for dealing with missing data in clinical trials: from design to analysis." *The Yale Journal of Biology and Medicine* 86.3 (2013): 343.

[64] "Missing Completely At Random (MCAR)." http://missingdata.lshtm.ac.uk/missing-completely-at-random-mcar 14th April 2014

[65] Brown, Helen, and Robin Prescott. "Applied mixed models in medicine." John Wiley & Sons, 2015.

[66] "Missing At Random (MAR)." http://missingdata.lshtm.ac.uk/missing-at-random-mar. 14th April 2014

[67] "Missing Not At Random (MNAR)." http://missingdata.lshtm.ac.uk/missing-not-at-random-mnar. 14th April 2014

[68] Daniels, Michael J., and Joseph W. Hogan. "Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis." *CRC Press*, 2008.

[69] Musil, Carol M., et al. "A comparison of imputation techniques for handling missing data." *Western Journal of Nursing Research* 24.7 (2002): 815-829.

[70] Wood, Angela M., Ian R. White, and Simon G. Thompson. "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals." *Clinical Trials* 1.4 (2004): 368-376.

[71] Acuna, Edgar, and Caroline Rodriguez : "The treatment of missing values and its effect on classifier accuracy. Classification, Clustering, and Data Mining Applications." *Springer Berlin Heidelberg* (2004). 639-647

[72] Alghannam, Abdullah F., et al. "Post-Exercise Protein Trial: Interactions between Diet and Exercise (PEPTIDE): study protocol for randomized controlled trial." *Trials* 15.1 (2014): 459.

[73] Desai, Manisha, et al. "The use of complete-case and multiple imputation-based analyses in molecular epidemiology studies that assess interaction effects." *Epidemiologic Perspectives & Innovations* 8.1 (2011): 5.

[74] Davey, Adam. "Statistical power analysis with missing data: A structural equation modeling approach." Routledge (2009).

[75] Fitzmaurice, Garrett M., Nan M. Laird, and James H. Ware. "Applied longitudinal analysis." Vol. 998. *John Wiley & Sons*, 2012.

[76] "Joint Modelling in SAS." http://www.hsph.harvard.edu/fitzmaur/lda/Ch14-program.pdf. 3rd January 2016

[77] Molenberghs, Geert, et al. "Analyzing incomplete longitudinal clinical trial data." *Biostatistics* 5.3 (2004): 445-464.

[78] Wood, Angela M., et al. "Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes." *International Journal of Epidemiology* 34.1 (2005): 89-99.

[79] Higgins, Julian PT, Ian R. White, and Angela M. Wood. "Imputation methods for missing outcome data in meta-analysis of clinical trials." *Clinical Trials* 5.3 (2008): 225-239.

[80] Rubin, Donald B. "Multiple imputation for nonresponse in surveys." Vol. 81. *John Wiley & Sons*, 2004.

[81] Enders, Craig K. "Applied missing data analysis." Guilford Publications, 2010.

[82] Van Buuren, Stef, and Karin Oudshoorn. "Flexible multivariate imputation by MICE." Leiden, The Netherlands: TNO Prevention Center (1999).

[83] Heyting, A., J. T. B. M. Tolboom, and J. G. A. Essers. "Statistical handling of drop-outs in longitudinal clinical trials." *Statistics in Medicine* 11.16 (1992): 2043-2061.

[84] Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data." *Journal of the American Statistical Association* 90.429 (1995): 106-121.

[85] Meng, Xiao-Li. "Multiple-imputation inferences with uncongenial sources of input." *Statistical Science* (1994): 538-558.

[86] Royston, Patrick. "Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables." *Stata Journal* 9.3 (2009): 466.

[87] Verbeke, Geert, and Geert Molenberghs. "Linear mixed models for longitudinal data." *Springer Science & Business Media*, 2009.

[88] Rubin, D; "Inference and Missing Data." *Biometrika* 1976, 63 (3): 581-592.

[89] Wood AM, White, IR, Thompson SG : "Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals." *Clin Trials* 1.4 (2004): 368-376

[90] Sterne, Jonathan AC, Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter : "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." *BMJ: British Medical Journal* 338 (2009): b2393.

[91] Hedeker D, Gibbons RD : "Missing Data in Longitudinal Studies." *Longitudinal Data Analysis*, Wiley 2006, 279-312,

[92] Rubin D: "Multiple Imputation for Nonresponse in Surveys." Wiley 1987

[93] Schafer JL: "Analysis of incomplete multivariate data." *Chapman and Hall* 1997

[94] Van Buuren, Stef, and Karin Oudshoorn : "Flexible multivariate imputation by MICE." (1999): 1-20.

[95] Lin, Danyu Y., and Lee-Jen Wei. "The robust inference for the Cox proportional hazards model." *Journal of the American Statistical Association* 84.408 (1989): 1074-1078.

[96] McCleary L: "Using multiple imputation for analysis of incomplete data in clinical research." *Nurs Res.* 51.5 (2002):339-43.

[97] Schulz KF, Altman DG, Moher D. and the CONSORT Group : "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials." *BMC Medicine* 8.18 (2010)

[98] Power, Michael J., and C. Freeman. "A randomized controlled trial of IPT versus CBT in primary care: With some cautionary notes about handling missing values in clinical trials." *Clinical psychology & psychotherapy* 19.2 (2012): 159-169.

[99] Charan, Jaykaran, and Tamoghna Biswas. "How to calculate sample size for different study designs in medical research?." *Indian journal of psychological medicine* 35.2 (2013): 121.

[100] Schoenfeld, David A. "Sample-size formula for the proportional-hazards regression model." *Biometrics* (1983): 499-503.

[101] Therneau, Terry M., and Patricia M. Grambsch. "Modeling survival data: extending the Cox model. *Springer Science & Business Media*, 2000.

[102] Rao, C. Radhakrishna. "Tests of significance in multivariate analysis." *Biometrika* (1948): 58-79.

[103] Rao, C. Radhakrishna. "Linear statistical inference and its applications." *John Wiley & Sons.* (2009):22

[104] Hsieh, F. Y., and Philip W. Lavori. "Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates." *Controlled Clinical Trials* 21.6 (2000): 552-560.

[105] Snijders, Tom AB. "Power and sample size in multilevel linear models." *Encyclopedia of Statistics in Behavioral Science* 3 (2005): 1570?1573

[106] "Properties of the Discrete Uniform Distribution." http://www.milefoot.com/math/stat/pdfd-uniformdisc.htm 14th February 2015

[107] Karow, A., et al. "PANSS syndromes and quality of life in schizophrenia." *Psychopathology* 38.6 (2005): 320-326.

[108] Chappell, Lucy C., et al. "Effect of antioxidants on the occurrence of pre-eclampsia in women at increased risk: a randomised trial." *The Lancet* 354.9181 (1999): 810-816.

[109] Arinaitwe, Emmanuel, et al. "Artemether-lumefantrine versus dihydroartemisinin-piperaquine for falciparum malaria: a longitudinal, randomized trial in young Ugandan children." *Clinical Infectious Diseases* 49.11 (2009): 1629-1637.

[110] Hedeker, Donald, Robert D. Gibbons, and Christine Waternaux. "Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups." *Journal of Educational and Behavioral Statistics* 24.1 (1999): 70-93.

[111] Overall, John E., and Suzanne R. Doyle. "Estimating sample sizes for repeated measurement designs." *Controlled Clinical Trials* 15.2 (1994): 100-123.

[112] Bender, Ralf, Thomas Augustin, and Maria Blettner. "Generating survival times to simulate Cox proportional hazards models." *Statistics In Medicine* 24.11 (2005): 1713-1723.

[113] "Garg, Mohan L., B. Raja Rao, and Carol K. Redmond. "Maximum-likelihood estimation of the parameters of the Gompertz survival function." *Applied Statistics* (1970): 152-159.

[114] Auleley, Guy-Robert, et al. "The methods for handling missing data in clinical trials influence sample size requirements." *Journal of Clinical Epidemiology* 57.5 (2004): 447-453.

[115] Sweeting, Michael J., and Simon G. Thompson. "Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture." *Biometrical Journal* 53.5 (2011): 750-763.

[116] Li, Ning, et al. "Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt:PA stroke trial." *Statistics In Medicine* 29.5 (2010): 546-557.

[117] Santos Nobre, Juvencio, and Julio da Motta Singer. "Residual analysis for linear mixed models." *Biometrical Journal* 49.6 (2007): 863-875.

[118] Cox, David R., and E. Joyce Snell. "A general definition of residuals." *Journal of the Royal Statistical Society. Series B (Methodological)* (1968): 248-275.

[119] Dobson, Angela, and Robin Henderson. "Diagnostics for joint longitudinal and dropout time modeling." *Biometrics* 59.4 (2003): 741-751.

[120] Rizopoulos, Dimitris, Geert Verbeke, and Geert Molenberghs. "Multiple Imputation Based Residuals and Diagnostic Plots for Joint Models of Longitudinal and Survival Outcomes." *Biometrics* 66.1 (2010): 20-29.

[121] Rizopoulos, Dimitris. "JM: An R package for the joint modelling of longitudinal and time-to-event data." *Journal of Statistical Software* 35.9 (2010): 1-33.

[122] Diggle, Peter J. "Testing for random dropouts in repeated measurement data." *Biometrics* (1989): 1255-1258.

[123] De Maesschalck, Roy, Delphine Jouan-Rimbaud, and Dsir L. Massart. "The mahalanobis distance." *Chemometrics and intelligent laboratory systems* 50.1 (2000): 1-18.

[124] Williams, James D., et al. "On the distribution of Hotelling's T2 statistic based on the successive differences covariance matrix estimator." *Journal of Quality Technology* 38 (2006): 217-229.

[125] Verbeke, Geert, Geert Molenberghs, and Caroline Beunckens. "Formal and informal model selection with incomplete data." *Statistical Science* (2008): 201-218.

[126] Gelman, Andrew, et al. "Multiple imputation for model checking: completed data plots with missing and latent data." *Biometrics* 61.1 (2005): 74-85.

[127] Van Buuren, Stef, and C. G. M. Oudshoorn. "Multivariate imputation by chained equations." MICE V1. 0 user's manual. Leiden: TNO Preventie en Gezondheid (2000).

[128] Chib, Siddhartha, and Edward Greenberg. "Understanding the metropolis-hastings algorithm." *The American Statistician* 49.4 (1995): 327-335.

[129] Cleveland, William S., and Susan J. Devlin. "Locally weighted regression: an approach to regression analysis by local fitting." *Journal of the American Statistical Association* 83.403 (1988): 596-610.

[130] Cook, R. Dennis. "Assessment of local influence." *Journal of the Royal Statistical Society. Series B (Methodological)* (1986): 133-169.

[131] Kim, Choongrak. "Cook's distance in spline smoothing." *Statistics & Probability Letters* 31.2 (1996): 139-144.

[132] Condor Liverpool, http://condor.liv.ac.uk. 23rd September 2013

[133] Cox, David Roxbee, and David Victor Hinkley. "Theoretical statistics." *CRC Press.* 1979.

[134] Fukuoka, Masahiro, et al. "Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non-small-cell lung cancer in Asia (IPASS)." *Journal of Clinical Oncology* 29.21 (2011): 2866-2874.

[135] Sorensen, A. Gregory, et al. "A "vascular normalization index" as potential mechanistic biomarker to predict survival after a single dose of cediranib in recurrent glioblastoma patients." *Cancer Research* 69.13 (2009): 5296-5300.

[136] Turner, Bruce C., et al. "BAG-1: a novel biomarker predicting long-term survival in early-stage breast cancer." *Journal of Clinical Oncology* 19.4 (2001): 992-1000.

[137] Scher, H. I., et al. "Evaluation of circulating tumor cell (CTC) enumeration as an efficacy response biomarker of overall survival (OS) in metastatic castration-resistant prostate cancer (mCRPC): Planned final analysis (FA) of COU-AA-301, a randomized, double-blind, placebo-controlled, phase III study of abiraterone acetate (AA) plus low-dose prednisone (P) post docetaxel." *ASCO Annual Meeting Proceedings.* 29.18 (2011).

[138] Casamassima, Addolorata, et al. "C-reactive protein: a biomarker of survival in patients with metastatic renal cell carcinoma treated with subcutaneous interleukin-2 based immunotherapy." *The Journal of Urology* 173.1 (2005): 52-55.

[139] McFadden Jr, E. R. "Acute severe asthma." *American Journal of Respiratory and Critical Care Medicine* 168.7 (2003): 740-759.

[140] Rizopoulos, Dimitris, Geert Verbeke, and Geert Molenberghs. "Multiple Imputation Based Residuals and Diagnostic Plots for Joint Models of Longitudinal and Survival Outcomes." *Biometrics* 66.1 (2010): 20-29.

[141] Anota, Amlie, et al. "Time to health-related quality of life score deterioration as a modality of longitudinal analysis for health-related quality of life studies in oncology: do we need RECIST for quality of life to achieve standardization?." *Quality of Life Research* 24.1 (2013): 5-18.

[142] Solfrizzi, Vincenzo, et al. "Dietary intake of unsaturated fatty acids and age-related cognitive decline: a 8.5-year follow-up of the Italian Longitudinal Study on Aging." *Neurobiology of Ageing* 27.11 (2006): 1694-1704.

[143] Lim, Eric, et al. "Longitudinal study of the profile and predictors of left ventricular mass regression after stentless aortic valve replacement." *The Annals of Thoracic Surgery* 85.6 (2008): 2026-2029.

[144] Croquet, Vincent, et al. "Prothrombin index is an indirect marker of severe liver fibrosis." *European Journal of Gastroenterology & Hepatology* 14.10 (2002): 1133-1141.

[145] Fieuws, Steffen, and Geert Verbeke. "Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach." *Statistics In Medicine* 23.20 (2004): 3093-3104.

[146] Fieuws, Steffen, and Geert Verbeke. "Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles." *Biometrics* 62.2 (2006): 424-431.

[147] Verbeke, Geert, et al. "The analysis of multivariate longitudinal data: a review." *Statistical Methods in Medical Research* 23.1 (2014): 42-59.

Appendix

Publications

The work carried out in Chapter 3 was published in *Trials* with the following citation:

Powney, M., Williamson, P., Kirkham, J., & Kolamunnage-Dona, R. (2014). A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*, 15(1), 1-19.

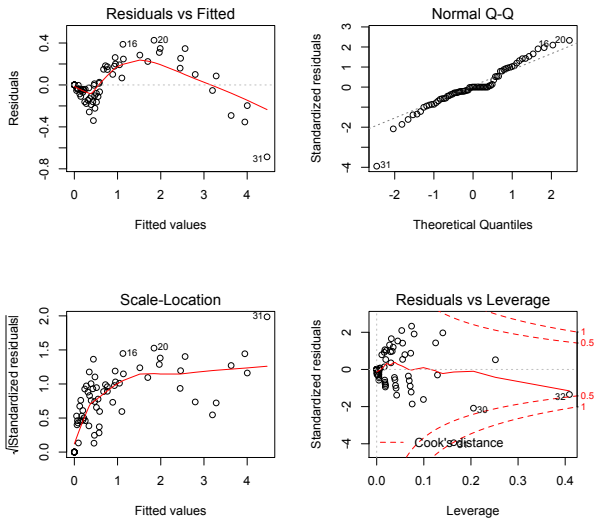
Extra Figures

Chapter 3 - Complete List of Medical Areas

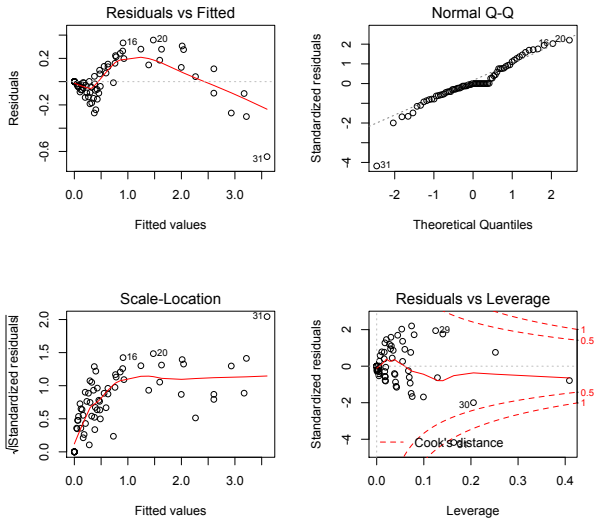
Medical area	Papers
Mental health	13
Cancer	11
Rheumatology	10
Infectious diseases	8
Heart and circulation	7
Dentistry/oral health	6
Neurology	6
Anaesthesia and pain control	6
Blood disorders	3
Developmental, psychosocial, and learning problems	2
Endocrine and metabolic	5
Eye and vision	2
Gastroenterology	1
Health care of older people	2
Kidney disease	2
Lungs and airways	2
Neonatal care	2
Orthopaedics and trauma	4
Pregnancy and childbirth	3
Skin	1
Urology	1
Wounds	3

Figure 8.1: Complete List of Medical Areas in Systematic Review

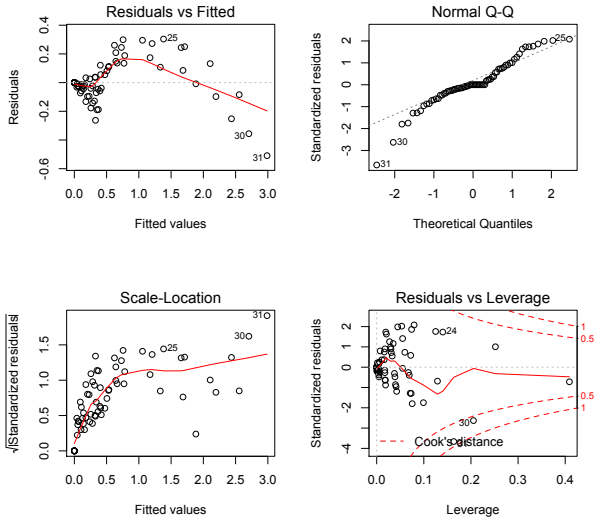
Chapter 5 - Diagnostic Plots of Model Fit



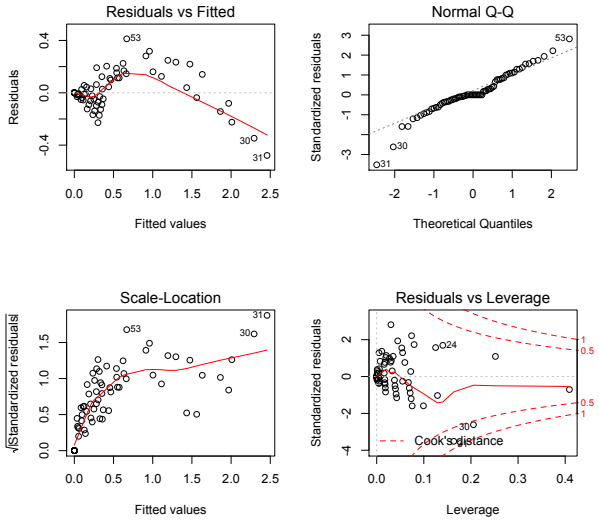
Diagnostic Plots for 10% Dropout Model



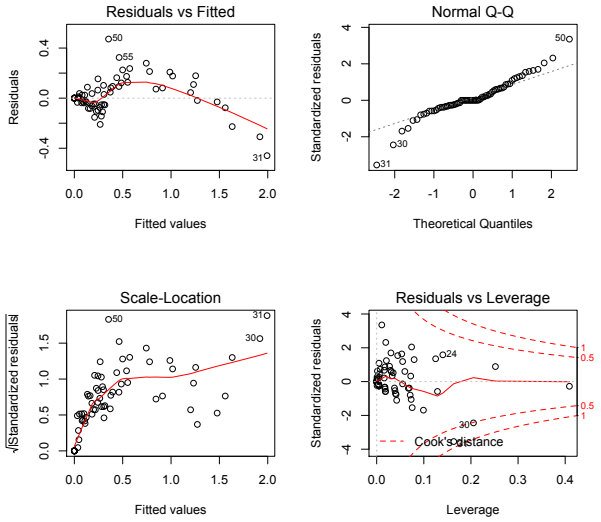
Diagnostic Plots for 20% Dropout Model



Diagnostic Plots for 30% Dropout Model



Diagnostic Plots for 40% Dropout Model



Diagnostic Plots for 50% Dropout Model

Complete List of Papers in the Systematic Review

1. Afshar H, Roohafza H, Mousavi G, Golchin S, Toghianifar N, Sadeghi M, Talaei M: "Topiramate add-on treatment in schizophrenia: a randomised, double-blind, placebo-controlled clinical trial". *Journal of psychopharmacology* 2009, 23(2): 157-162.
2. Ambrosino JM, Fennie K, Whittemore R, Jaser S, Dowd MF, Grey M: "Short-term effects of coping skills training in school-age children with type 1 diabetes". *Pediatric diabetes* 2008, 9(3 Pt 2):74-82.
3. Arteaga-Troncoso G, Villegas-Alvarado A, Belmont-Gomez A, Martinez-Herrera FJ, Villagrana-Zesati R, Guerra-Infante F: "Intracervical application of the nitric oxide donor isosorbide dinitrate for induction of cervical ripening: a randomised controlled trial to determine clinical efficacy and safety prior to first trimester surgical evacuation of retained products of conception". *BJOG : an international journal of obstetrics and gynaecology* 2005, 112(12):1615-1619.
4. Berger AM, Kuhn BR, Farr LA, Lynch JC, Agrawal S, Chamberlain J, Von Essen SG: "Behavioral therapy intervention trial to improve sleep quality and cancer-related fatigue". *Psycho-oncology* 2009, 18(6):634-646.
5. Bird SE, Williams K, Kula K: "Preoperative acetaminophen vs ibuprofen for control of pain after orthodontic separator placement". *American journal of orthodontics and dentofacial orthopedics : official publication of the American Association of Orthodontists, its constituent societies, and the American Board of Orthodontics* 2007, 132(4):504-510.
6. Callahan LF, Mielenz T, Freburger J, Shreffler J, Hootman J, Brady T, Buysse K, Schwartz T: "A randomized controlled trial of the people with arthritis can exercise program: symptoms, function, physical activity, and psychosocial outcomes". *Arthritis and rheumatism* 2008, 59(1):92-101.
7. Chae J, Yu DT, Walker ME, Kirsteins A, Elovic EP, Flanagan SR, Harvey RL, Zorowitz RD, Frost FS, Grill JH *et al*: "Intramuscular Electrical Stimulation for Hemiplegic Shoulder Pain". *American Journal of Physical Medicine & Rehabilitation* 2005, 84(11):832-842.
8. Cortelli JR, Querido SM, Aquino DR, Ricardo LH, Pallos D: "Longitudinal clinical evaluation of adjunct minocycline in the treatment of chronic periodontitis". *Journal of periodontology* 2006, 77(2):161-166.
9. Doheny MO, Sedlak CA, Hall RJ, Estok PJ: "Structural model for osteoporosis preventing behavior in men". *American journal of men's health* 2010, 4(4): 334-343.
10. Dorstyn DS, Mathias JL, Denson LA: "Psychological intervention during spinal rehabilitation: a preliminary study". *Spinal cord* 2010, 48(10):756-761.
11. Dougherty CM, Thompson EA, Lewis FM: "Long-term outcomes of a telephone intervention after an ICD". *Pacing and clinical electrophysiology : PACE* 2005, 28(11):1157-1167.
12. Elnour AA, El Mugammar IT, Jaber T, Revel T, McElnay JC: "Pharmaceutical care of patients with gestational diabetes mellitus". *Journal of evaluation in clinical practice* 2008, 14(1):131-140.
13. Estok PJ, Sedlak CA, Doheny MO, Hall R: "Structural model for osteoporosis preventing behavior in postmenopausal women". *Nursing research* 2007, 56(3):

148-158.
14. Fader JP, Cleary RK, Lampman RM, Winter S, Singal BM, Plona AE: “Does intrathecal morphine sulfate provide preemptive analgesia for patients undergoing stapled hemorrhoidopexy?” *Pain medicine (Malden, Mass)* 2011, 12(2):322-327.
15. Fernandez WG, Winter MR, Mitchell PM, Bullock H, Donovan J, St George J, Feldman JA, Gallagher SS, McKay MP, Bernstein E *et al*: “Six-month follow-up of a brief intervention on self-reported safety belt use among emergency department patients”. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine* 2009, 16(11):1221-1224.
16. Fields C, Drye L, Vaidya V, Lyketsos C, Group AR: “Celecoxib or naproxen treatment does not benefit depressive symptoms in persons age 70 and older: findings from a randomized controlled trial”. *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry* 2012, 20(6):505-513.
17. Fortin PR, Abrahamowicz M, Ferland D, Lacaille D, Smith CD, Zummer M, Canadian Network For Improved Outcomes in Systemic L: “Steroid-sparing effects of methotrexate in systemic lupus erythematosus: a double-blind, randomized, placebo-controlled trial.” *Arthritis and rheumatism* 2008, 59(12): 1796-1804.
18. Frawley HC, Phillips BA, Bo K, Galea MP: “Physiotherapy as an adjunct to prolapse surgery: an assessor-blinded randomized controlled trial.” *Neurourology and urodynamics* 2010, 29(5):719-725.
19. Friese RS, Barber R, McBride D, Bender J, Gentilello LM: “Could Beta blockade improve outcome after injury by modulating inflammatory profiles?” *The Journal of trauma* 2008, 64(4):1061-1068.
20. Fukui S, Ogawa K, Yamagishi A: “Effectiveness of communication skills training of nurses on the quality of life and satisfaction with healthcare professionals among newly diagnosed cancer patients: a preliminary study”. *Psycho-oncology* 2011, 20(12):1285-1291.
21. Garrett WE, Kaeding CC, ElAttrache NS, Xerogeanes JW, Hewitt MS, Skrepnik NV, Papilion JD, O'Donnell JB, Fox DL, Ruvuna F *et al*: “Novel drug OMS103HP reduces pain and improves joint motion and function for 90 days after arthroscopic meniscectomy.” *Arthroscopy : the journal of arthroscopic & related surgery : official publication of the Arthroscopy Association of North America and the International Arthroscopy Association* 2011, 27(8):1060-1070.
22. Goldman RH, Stason WB, Park SK, Kim R, Mudgal S, Davis RB, Kaptchuk TJ: “Low-dose amitriptyline for treatment of persistent arm pain due to repetitive use”. *Pain* 2010, 149(1):117-123.
23. Goodyer I, Dubicka B, Wilkinson P, Kelvin R, Roberts C, Byford S, Breen S, Ford C, Barrett B, Leech A *et al*: “Selective serotonin reuptake inhibitors (SSRIs) and routine specialist care with and without cognitive behaviour therapy in adolescents with major depression: randomised controlled trial.” *Bmj* 2007, 335(7611):142.
24. Grandes G, Sanchez A, Montoya I, Ortega Sanchez-Pinilla R, Torcal J, Group P: “Two-year longitudinal analysis of a cluster randomized trial of physical activity

promotion by general practitioners". *PloS one* 2011, 6(3):e18363.
25. Gustavsson C, Denison E, von Koch L: "Self-management of persistent neck pain: two-year follow-up of a randomized controlled trial of a multicomponent group intervention in primary health care." *Spine* 2011, 36(25):2105-2115.
26. Hansen NB, Tarakeshwar N, Ghebremichael M, Zhang H, Kochman A, Sikkema KJ: "Longitudinal effects of coping on outcome in a randomized controlled trial of a group intervention for HIV-positive adults with AIDS-related bereavement." *Death studies* 2006, 30(7):609-636.
27. Harrington JW, Logan S, Harwell C, Gardner J, Swingle J, McGuire E, Santos R: "Effective analgesia using physical interventions for infant immunizations". *Pediatrics* 2012, 129(5):815-822.
28. Hatfield LA: "Sucrose decreases infant biobehavioral pain response to immunizations: a randomized controlled trial". *Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing / Sigma Theta Tau* 2008, 40(3):219-225.
29. Hatfield LA, Gusic ME, Dyer AM, Polomano RC: "Analgesic properties of oral sucrose during routine immunizations at 2 and 4 months of age". *Pediatrics* 2008, 121(2):e327-334.
30. Heiney SP, Millon Underwood S, Tavakoli A, Arp Adams S, Wells LM, Bryant LH: "Randomized trial of therapeutic group by teleconference: African American women with breast cancer." *Cancer* 2012, 118(15):3822-3832.
31. Holmes-Rovner M, Stommel M, Corser WD, Olomu A, Holtrop JS, Siddiqi A, Dunn SL: "Does outpatient telephone coaching add to hospital quality improvement following hospitalization for acute coronary syndrome?" *Journal of general internal medicine* 2008, 23(9):1464-1470.
32. Hovell MF, Nichols JF, Irvin VL, Schmitz KE, Rock CL, Hofstetter CR, Keating K, Stark LJ: "Parent/Child training to increase preteens' calcium, physical activity, and bone density: a controlled trial." *American journal of health promotion : AJHP* 2009, 24(2):118-128.
33. Howell EA, Balbierz A, Wang J, Parides M, Zlotnick C, Leventhal H: "Reducing postpartum depressive symptoms among black and Latina mothers: a randomized controlled trial." *Obstetrics and gynecology* 2012, 119(5):942-949.
34. Husebo BS, Ballard C, Sandvik R, Nilsen OB, Aarsland D: "Efficacy of treating pain to reduce behavioural disturbances in residents of nursing homes with dementia: cluster randomised clinical trial." *Bmj* 2011, 343:d4065.
35. Iannotti LL, Zavaleta N, Leon Z, Shankar AH, Caulfield LE: "Maternal zinc supplementation and growth in Peruvian infants." *The American journal of clinical nutrition* 2008, 88(1):154-160.
36. James J, Thomas P, Kerr D: "Preventing childhood obesity: two year follow-up results from the Christchurch obesity prevention programme in schools (CHOPPS)". *Bmj* 2007, 335(7623):762.
37. Jenkins AL, Kacinik V, Lyon M, Wolever TM: "Effect of adding the novel fiber, PGX(R), to commonly consumed foods on glycemic response, glycemic index and GRIP: a simple and effective strategy for reducing post prandial blood glucose levels--a randomized, controlled trial." *Nutrition journal* 2010, 9:58.
38. Kaisar MO, Wiggins KJ, Sturtevant JM, Hawley CM, Campbell SB, Isbel NM,

Mudge DW, Bofinger A, Petrie JJ, Johnson DW: “A randomized controlled trial of fludrocortisone for the treatment of hyperkalemia in hemodialysis patients.” *American journal of kidney diseases : the official journal of the National Kidney Foundation* 2006, 47(5):809-814.

39. Kaplan K, Salzer MS, Solomon P, Brusilovskiy E, Cousounis P: “Internet peer support for individuals with psychiatric disabilities: A randomized controlled trial”. *Social science & medicine* 2011, 72(1):54-62.
40. Katrak S, Gasasira A, Arinaitwe E, Kakuru A, Wanzira H, Bigira V, Sandison TG, Homsy J, Tappero JW, Kamya MR *et al*: “Safety and tolerability of artemether-lumefantrine versus dihydroartemisinin-piperaquine for malaria in young HIV-infected and uninfected children”. *Malaria journal* 2009, 8:272.
41. Keefe MR, Kajrlsen KA, Lobo ML, Kotzer AM, Dudley WN: “Reducing parenting stress in families with irritable infants.” *Nursing research* 2006, 55(3):198-205.
42. Kendall E, Catalano T, Kuipers P, Posner N, Buys N, Charker J: “Recovery following stroke: the role of self-management education.” *Social science & medicine* 2007, 64(3):735-746.
43. Klehmet J, Harms H, Richter M, Prass K, Volk HD, Dirnagl U, Meisel A, Meisel C: “Stroke-induced immunodepression and post-stroke infections: lessons from the preventive antibacterial therapy in stroke trial”. *Neuroscience* 2009, 158(3): 1184-1193.
44. Kolls BJ, Stacy M: “Apomorphine”. *Clinical Neuropharmacology* 2006, 29(5): 292-301.
45. Kravitz RL, Tancredi DJ, Jerant A, Saito N, Street RL, Grennan T, Franks P: “Influence of patient coaching on analgesic treatment adjustment: secondary analysis of a randomized controlled trial.” *Journal of pain and symptom management* 2012, 43(5):874-884.
46. Kuhnle MD, Ryan DS, Coe CD, Eaddy J, Kuzmowych C, Edwards J, Howard RS, Bower KS: “Oral gabapentin for photorefractive keratectomy pain.” *Journal of cataract and refractive surgery* 2011, 37(2):364-369.
47. Lai HL, Hwang MJ, Chen CJ, Chang KF, Peng TC, Chang FM: “Randomised controlled trial of music on state anxiety and physiological indices in patients undergoing root canal treatment.” *Journal of clinical nursing* 2008, 17(19): 2654-2660.
48. Langhammer B, Stanghelle JK, Lindmark B: “An evaluation of two different exercise regimes during the first year following stroke: a randomised controlled trial.” *Physiotherapy theory and practice* 2009, 25(2):55-68.
49. Lara MA, Navarro C, Navarrete L: “Outcome results of a psycho-educational intervention in pregnancy to prevent PPD: a randomized control trial.” *Journal of affective disorders* 2010, 122(1-2):109-117.
50. Lee JT, Yen HW: “Randomized controlled evaluation of a theory-based postpartum sexual health education programme.” *Journal of advanced nursing* 2007, 60(4):389-401.
51. Liedl A, Muller J, Morina N, Karl A, Denke C, Knaevelsrud C: “Physical activity within a CBT intervention improves coping with pain in traumatized refugees: results of a randomized controlled design”. *Pain medicine (Malden, Mass)* 2011, 12(2):234-245.

52. Lin PH, Appel LJ, Funk K, Craddick S, Chen C, Elmer P, McBurnie MA, Champagne C: “The PREMIER intervention helps participants follow the Dietary Approaches to Stop Hypertension dietary pattern and the current Dietary Reference Intakes recommendations”. *Journal of the American Dietetic Association* 2007, 107(9):1541-1551.
53. Lo C, Burman D, Hales S, Swami N, Rodin G, Zimmermann C: “The FAMCARE-Patient scale: measuring satisfaction with care of outpatients with advanced cancer.” *European journal of cancer* 2009, 45(18):3182-3188.
54. Lorenzo R, Garcia V, Orsini M, Martin C, Sanz M: “Clinical efficacy of a xenogeneic collagen matrix in augmenting keratinized mucosa around implants: a randomized controlled prospective clinical trial.” *Clinical oral implants research* 2012, 23(3):316-324.
55. Lovell DJ, Glass D, Ranz J, Kramer S, Huang B, Sierra RI, Henderson CJ, Passo M, Graham B, Bowyer S *et al*: “A randomized controlled trial of calcium supplementation to increase bone mineral density in children with juvenile rheumatoid arthritis.” *Arthritis and rheumatism* 2006, 54(7):2235-2242.
56. Marcantonio ER, Palihnich K, Appleton P, Davis RB: “Pilot randomized trial of donepezil hydrochloride for delirium after hip fracture.” *Journal of the American Geriatrics Society* 2011, 59 Suppl 2:S282-288.
57. Markland AD, Richter HE, Burgio KL, Myers DL, Hernandez AL, Subak LL: “Weight loss improves fecal incontinence severity in overweight and obese women with urinary incontinence”. *International urogynecology journal* 2011, 22(9):1151-1157.
58. Marr M, Baker J, Lambon N, Perry J: “The effects of the Bowen technique on hamstring flexibility over time: a randomised controlled trial.” *Journal of bodywork and movement therapies* 2011, 15(3):281-290.
59. Maturi MS, Afshary P, Abedi P: “Effect of physical activity intervention based on a pedometer on physical activity level and anthropometric measures after childbirth: a randomized controlled trial.” *BMC pregnancy and childbirth* 2011, 11:103.
60. McCurry SM, Gibbons LE, Logsdon RG, Vitiello MV, Teri L: “Nighttime insomnia treatment and education for Alzheimer's disease: a randomized, controlled trial”. *Journal of the American Geriatrics Society* 2005, 53(5):793-802.
61. McKinley S, Dracup K, Moser DK, Riegel B, Doering LV, Meischke H, Aitken LM, Buckley T, Marshall A, Pelter M: “The effect of a short one-on-one nursing intervention on knowledge, attitudes and beliefs related to response to acute coronary syndrome in people with coronary heart disease: a randomized controlled trial.” *International journal of nursing studies* 2009, 46(8):1037-1046.
62. McMurray JS, Connor N, Ford CN: “Cidofovir efficacy in recurrent respiratory papillomatosis: a randomized, double-blind, placebo-controlled study.” *The Annals of otology, rhinology, and laryngology* 2008, 117(7):477-483.
63. Meneses K, McNees P, Azuero A, Loerzel VW, Su X, Hassey LA: “Preliminary evaluation of psychoeducational support interventions on quality of life in rural breast cancer survivors after primary treatment.” *Cancer nursing* 2009, 32(5): 385-397.
64. Miyashita M: “Effects of continuous versus accumulated activity patterns on

postprandial triacylglycerol concentrations in obese men." *International journal of obesity* 2008, 32(8):1271-1278.
65. Naar-King S, Parsons JT, Murphy DA, Chen X, Harris DR, Belzer ME: "Improving health outcomes for youth living with the human immunodeficiency virus: a multisite randomized trial of a motivational intervention targeting multiple risk behaviors." *Archives of pediatrics & adolescent medicine* 2009, 163(12): 1092-1098.
66. Nabulsi MM, Tamim H, Mahfoud Z, Itani M, Sabra R, Chamseddine F, Mikati M: "Alternating ibuprofen and acetaminophen in the treatment of febrile children: a pilot study [ISRCTN30487061]". *BMC medicine* 2006, 4:4.
67. Neiva RF, Al-Shammari K, Nociti FH, Jr., Soehren S, Wang HL: "Effects of vitamin-B complex supplementation on periodontal wound healing." *Journal of periodontology* 2005, 76(7):1084-1091.
68. Orrell M, Spector A, Thorgrimsen L, Woods B: "A pilot study examining the effectiveness of maintenance Cognitive Stimulation Therapy (MCST) for people with dementia." *International journal of geriatric psychiatry* 2005, 20(5):446-451.
69. Pijls BG, Valstar ER, Kaptein BL, Fiocco M, Nelissen RG: "The beneficial effect of hydroxyapatite lasts: a randomized radiostereometric trial comparing hydroxyapatite-coated, uncoated, and cemented tibial components for up to 16 years." *Acta orthopaedica* 2012, 83(2):135-141.
70. Poole H, Glenn S, Murphy P: "A randomised controlled study of reflexology for the management of chronic low back pain." *European journal of pain* 2007, 11(8): 878-887.
71. Potterton J, Stewart A, Cooper P, Becker P: "The effect of a basic home stimulation programme on the development of young children infected with HIV." *Developmental medicine and child neurology* 2010, 52(6):547-551.
72. Power MJ, Freeman C: "A randomized controlled trial of IPT versus CBT in primary care: with some cautionary notes about handling missing values in clinical trials." *Clinical psychology & psychotherapy* 2012, 19(2):159-169.
73. Pradhan EK, Baumgarten M, Langenberg P, Handwerger B, Gilpin AK, Magyari T, Hochberg MC, Berman BM: "Effect of Mindfulness-Based Stress Reduction in rheumatoid arthritis patients." *Arthritis and rheumatism* 2007, 57(7):1134-1142.
74. Pritchett Y, Jemiai Y, Chang Y, Bhan I, Agarwal R, Zoccali C, Wanner C, Lloyd-Jones D, Cannata-Andia JB, Thompson T *et al*: "The use of group sequential, information-based sample size re-estimation in the design of the PRIMO study of chronic kidney disease." *Clinical trials* 2011, 8(2):165-174.
75. Pulido JS, Winters JL, Boyer D: "Preliminary analysis of the final multicenter investigation of rheopheresis for age related macular degeneration (AMD) trial (MIRA-1) results." *Transactions of the American Ophthalmological Society* 2006, 104:221-231.
76. Richardson LC, Wang W, Hartzema AG, Wagner S: "The role of health-related quality of life in early discontinuation of chemotherapy for breast cancer." *The breast journal* 2007, 13(6):581-587.
77. Robinson P, Keenan AM, Conaghan PG: "Clinical effectiveness and dose response of image-guided intra-articular corticosteroid injection for hip osteoarthritis." *Rheumatology* 2007, 46(2):285-291.

78. Romieu I, Riojas-Rodriguez H, Marron-Mares AT, Schilmann A, Perez-Padilla R, Masera O: “Improved biomass stove intervention in rural Mexico: impact on the respiratory health of women.” *American journal of respiratory and critical care medicine* 2009, 180(7):649-656.
79. Rostami HR, Malamiri RA: “Effect of treatment environment on modified constraint-induced movement therapy results in children with spastic hemiplegic cerebral palsy: a randomized controlled trial.” *Disability and rehabilitation* 2012, 34(1):40-44.
80. Sadik A, Yousif M, McElnay JC: “Pharmaceutical care of patients with heart failure.” *British journal of clinical pharmacology* 2005, 60(2):183-193.
81. Sanz M, Lorenzo R, Aranda JJ, Martin C, Orsini M: “Clinical evaluation of a new collagen matrix (Mucograft prototype) to enhance the width of keratinized tissue in patients with fixed prosthetic restorations: a randomized prospective clinical trial.” *Journal of clinical periodontology* 2009, 36(10):868-876.
82. Shiga T, Yamada Y, Matsuda N, Tanaka T, Urae A, Hashiguchi M, Hagiwara N, Kasanuki H: “Influence of cilnidipine or nisoldipine on sympathetic activity in healthy male subjects”. *Heart and vessels* 2007, 22(6):404-409.
83. Shim JC, Jung DU, Jung SS, Seo YS, Cho DM, Lee JH, Lee SW, Kong BG, Kang JW, Oh MK *et al*: “Adjunctive varenicline treatment with antipsychotic medications for cognitive impairments in people with schizophrenia: a randomized double-blind placebo-controlled trial.” *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 2012, 37(3):660-668.
84. Shiroiwa T, Fukuda T, Shimozuma K, Kuranami M, Suemasu K, Ohashi Y, Watanabe T: “Comparison of EQ-5D scores among anthracycline-containing regimens followed by taxane and taxane-only regimens for node-positive breast cancer patients after surgery: the N-SAS BC 02 trial.” *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2011, 14(5):746-751.
85. Song L, Northouse LL, Braun TM, Zhang L, Cimprich B, Ronis DL, Mood DW: “Assessing longitudinal quality of life in prostate cancer patients and their spouses: a multilevel modeling approach.” *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2011, 20(3):371-381.
86. Song L, Northouse LL, Zhang L, Braun TM, Cimprich B, Ronis DL, Mood DW: “Study of dyadic communication in couples managing prostate cancer: a longitudinal perspective.” *Psycho-oncology* 2012, 21(1):72-81.
87. Stanton B, Guo J, Cottrell L, Galbraith J, Li X, Gibson C, Pack R, Cole M, Marshall S, Harris C: “The complex business of adapting effective interventions to new populations: an urban to rural transfer.” *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 2005, 37(2):163.
88. Sun JL, Sung MS, Huang MY, Cheng GC, Lin CC: “Effectiveness of acupressure for residents of long-term care facilities with insomnia: a randomized controlled trial.” *International journal of nursing studies* 2010, 47(7):798-805.
89. Swanenburg J, de Bruin ED, Stauffacher M, Mulder T, Uebelhart D: “Effects of exercise and nutrition on postural balance and risk of falling in elderly people with

decreased bone mineral density: randomized controlled trial pilot study." *Clinical rehabilitation* 2007, 21(6):523-534.

90. Thorsdottir I, Tomasson H, Gunnarsdottir I, Gisladottir E, Kiely M, Parra MD, Bandarra NM, Schaafsma G, Martinez JA: "Randomized trial of weight-loss-diets for young adults varying in fish and fish oil content." *International journal of obesity* 2007, 31(10):1560-1566.
91. Tiersky LA, Anselmi V, Johnston MV, Kurtyka J, Roosen E, Schwartz T, Deluca J: "A trial of neuropsychologic rehabilitation in mild-spectrum traumatic brain injury." *Archives of physical medicine and rehabilitation* 2005, 86(8):1565-1574.
92. Trappenburg JC, Monninkhof EM, Bourbeau J, Troosters T, Schrijvers AJ, Verheij TJ, Lammers JW: "Effect of an action plan with ongoing support by a case manager on exacerbation-related outcome in patients with COPD: a multicentre randomised controlled trial." *Thorax* 2011, 66(11):977-984.
93. Tsui JI, Herman DS, Kettavong M, Anderson BJ, Stein MD: "Escitalopram is associated with reductions in pain severity and pain interference in opioid dependent patients with depressive symptoms." *Pain* 2011, 152(11):2640-2644.
94. Urech C, Fink NS, Hoesli I, Wilhelm FH, Bitzer J, Alder J: "Effects of relaxation on psychobiological wellbeing during pregnancy: a randomized controlled trial". *Psychoneuroendocrinology* 2010, 35(9):1348-1355.
95. Vempati R, Bijlani RL, Deepak KK: "The efficacy of a comprehensive lifestyle modification programme based on yoga in the management of bronchial asthma: a randomized controlled trial." *BMC pulmonary medicine* 2009, 9:37.
96. Venzke L, Calvert JF, Jr., Gilbertson B: "A randomized trial of acupuncture for vasomotor symptoms in post-menopausal women." *Complementary therapies in medicine* 2010, 18(2):59-66.
97. Vigano A, Giacomet V, Pariani E, Giani E, Manfredini V, Bedogni G, Erba P, Amendola A, Zanetti A, Zuccotti G: "Long-term immunogenicity after one and two doses of a monovalent MF59-adjuvanted A/H1N1 Influenza virus vaccine coadministered with the seasonal 2009-2010 nonadjuvanted Influenza virus vaccine in HIV-infected children, adolescents, and young adults in a randomized controlled trial." *Clinical and vaccine immunology : CVI* 2011, 18(9):1503-1509.
98. Wang TJ, Belza B, Elaine Thompson F, Whitney JD, Bennett K: "Effects of aquatic exercise on flexibility, strength and aerobic fitness in adults with osteoarthritis of the hip or knee." *Journal of advanced nursing* 2007, 57(2): 141-152.
99. Whiteley P, Haracopos D, Knivsberg AM, Reichelt KL, Parlar S, Jacobsen J, Seim A, Pedersen L, Schondel M, Shattock P: "The ScanBrit randomised, controlled, single-blind study of a gluten- and casein-free dietary intervention for children with autism spectrum disorders." *Nutritional neuroscience* 2010, 13(2): 87-100.
100. Yang XL, Li HH, Hong MH, Kao HS: "The effects of Chinese calligraphy handwriting and relaxation training in Chinese Nasopharyngeal Carcinoma patients: a randomized controlled trial". *International journal of nursing studies* 2010, 47(5):550-559.

Novel Code

Chapter 2

```
#The following function imputes High/Low values for missing longitudinal data based on MAGNETIC
     for a dataset in wide format.


highlow<-function(set){
        subj=dim(set)[1]
        vars=dim(set)[2]

        for (i in 1:subj){
                        for (j in 2:vars){

                        if(is.na(set[i,j]) == TRUE && set[i,j-1] > 4.5){
                                        set[i,j] = 9
                                                        }


                        if(is.na(set[i,j]) == TRUE && set[i,j-1] <= 4.5){

                                set[i,j] = 0

                        }
                }
        }
set
}
```

Chapter 5

```
#The following function is used to simulate data for simulation study 2. This is a modified
     version of the simjoint function in the joineR package such that all patients prior to the
     end of follow up experienced an event, and post follow up these were censored. As we were
     interested in the variances of u1 and u2, but not the covariance - this was set to 0.

`simjoint` <-
function(n = 500, model = c("intslope", "int", "quad"), sepassoc = FALSE,
ntms = 5, b1 = c(0.05, 0.05, 0.03), b2 = c(1), gamma = c(0.04, 0.1), sigu,
vare = 0.5, theta0 = -3, theta1 = 1, censoring = TRUE, censlam = exp(-3),
truncation = FALSE, trunctime = max(ntms), gridstep = 0.01,u1=1.19,u2=0.00003)
{

model <- match.arg(model)

sigu=matrix(c(u1,0,0,u2),nrow=2,ncol=2)

#If a model is of an unrecognised type then give then print "Unknown model"
if(model!="intslope"&&model!="int"&&model!="quad")
```

```
{stop(paste("Unknown model", model))}

#If int model, set ran to 1, for intslope set ran to 2, if quad model set ran to 3
ran=2
if(model=="int"){ran=1}
if(model=="quad"){ran=3}

#Set lat to ran
lat=ran

#If sepassoc is false then set lat = 1
if(!sepassoc){lat=1}

#if the number of association parameters dont match the model choice then print an error message
if(length(gamma)!=lat){
cat("Warning: Number of association parameters do not match model choice\n")}

#set gamma = gamma repeated with the number of times being dependent on the model choice
gamma=rep(gamma,length=ran)

#If sigu is not specified in the functional call then set sigu to be a matrix with one element
    being equal to ran
if(missing(sigu)){sigu<-diag(ran)}

#If the coviariance matrix does not have the same number of horizontal elements as the model
    choice squared then flag up as an error.
if(length(sigu)!=ran^2){
#cat("Warning: Dimension of covariance matrix does not match chosen model\n")

if(length(sigu)>ran^2){sigu=sigu[1:ran,1:ran]}
else{sigu=diag(ran)*sigu[1]}}
#If variance sigu is less than 0 then give error message "Variance must be positive
if(model=="int"){if(sigu<0){stop("Variance must be positive")}}

#Error message appears if sigu is not symmetric
else{if(!isSymmetric(sigu)){
stop("Covariance matrix is not symmetric")}

#if the eigenvalues or the determinant of the covariance matrix are negative or if the
    determinant is zero then it is not positive semi-definite, so return an error message.
if(any(eigen(sigu)$values<0)||(det(sigu)<=0)){
stop("Covariance matrix must be positive semi-definite")}}

#The following function
"getD1"<-function(q,arg)
{
D<-matrix(0,q,length(arg))
for(i in 1:q){D[i,]=arg^(i-1)}
D
}


"simdat"<-function(n,model,sepassoc,ntms,b1,b2,gamma,sigu,vare,theta0,theta1,censoring,censlam,
    truncation,trunctime,gridstep)
```

```
{

#Binarise function is used to set the treatment groups to either 0 or 1.

binarise<-function(a){

if(a<0){

a=0

}

else { a=1

}
}


#Generate n values for(i in 1:n){
binx<-runif(n,-sqrt(3),sqrt(3))

for(i in 1:n) {
        binx[i] = binarise(binx[i])


}
#Combine these values in a data frame X2
X2<-cbind(binx)
#Give each individual values a unique id number
id<-1:n

#repeat this dependent on the number of times value inputted in the function, repeat for the
    other created values.
idl<-rep(id,each=ntms)
binxl<-rep(binx,each=ntms)

#set time values ranging from 0 to ntms-1, which creates n*ntms entries
time<-rep(seq(0,ntms-1,by=1),length=n*ntms)

#combine these into a data set which includes an intercept of 1 for all values and the time
    function
X1<-cbind(intercept=1,binxl,ltime=time)


#Generate n values U from the multvariate normal distribution with mu being alternating from 0 to
     ran (which is based on model choice), and variance sigu.
U<-mvrnorm(n,mu=rep(0,ran),Sigma=sigu)

#Put all details of U into a separate function Ul ntms times
Ul<-U[rep(1:n,each=ntms),]

#Generate matrix from the function getD1 above based on model choice
D<-getD1(ran,time)
```

```
#Create Matrix DU of time values multiplied by random generated values from mvrnorm
DU<-t(D)*U1
#This bit was rounded in the joint modelling analysis
Y<-X1%*%b1+rowSums(DU)+sqrt(vare)*rnorm(n*ntms)

sk=dim(Y)[1]

#Intercept values are extracted
u0<-U[,1]
#if slope model chosen then slope values are extracted
if(model=="intslope"){u1=U[,2]}
else{u1=rep(0,n)}
#The following calculates predictive values

b2x<-X2%*%b2



cens<-rep(1,n)
if(!sepassoc){gamma=rep(gamma[1],ran)}
if(model!="quad"){
if(model=="int"){gamma=c(gamma[1],0)}
uu<-runif(n)

#In each model, the survival times are simply generated from the Cox Proportional Hazards model
     with Gompertz baseline using the following code.
if(model=="int")
{
survtime<--log(uu)/exp(theta0+b2x+gamma[1]*u0)
}
else{
ii<-1*((theta1+gamma[2]*u1)<0)&(uu<exp(exp(theta0+b2x+gamma[1]*u0)/(theta1+gamma[2]*u1)))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-1*(log(1-(theta1+gamma[2]*u1[!ii])*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*u0[!
     ii]))/(theta1+gamma[2]*u1[!ii]))
}}
else{
tau<-trunctime
tgrid<-seq(runif(1,0,gridstep),tau,gridstep)
lam0<-exp(theta0+theta1*tgrid)
hazt<-gridstep*exp(b2x)%*%lam0
gD2<-gamma*getD1(ran,tgrid)
hmat<-exp(U%*%gD2)*hazt
uu<-matrix(runif(length(hmat)),n,length(tgrid))
tmat<-matrix(tgrid,n,length(tgrid),byrow=T)
tmat[hmat<uu]<-tau
survtime<-apply(tmat,1,min)
cens[survtime==tau]=0}

#If there are censored values, then generate them randomly as a function of the uniform
     distribution with parameter n.

if(censoring){censtime=-log(runif(n))/censlam}
```

```
#Otherwise, set the censored time as a very small time after the final time point (0.000000001),
     as the first time point is at t=0.
else{censtime=rep(ntms-0.999999,n)}

#If using an intercept or intercept with slope model with trancation then censored time is the
     minimum of the previously defined censored time and truncation time.
if(model!="quad"){if(truncation){censtime=pmin(censtime,trunctime)}}


#The following code generates a percentage of individuals who experienced an event/dropout.
ii<-censtime<survtime
survtime[ii]<-censtime[ii]
cens[ii]<-0
ls<-rep(survtime,each=ntms)
Y<-Y[ls>time]
X1<-X1[ls>time,]
idl<-idl[ls>time]
time<-time[ls>time]
cat(100*sum(cens)/n,"% experienced event\n")

list(longdat=data.frame(id=idl,Y,time,X1),survdat=data.frame(id,survtime,cens,X2))
}



sim<-simdat(n,model,sepassoc,ntms,b1,b2,gamma,sigu,vare,theta0,theta1,censoring,censlam,
     truncation,trunctime,gridstep)
list(longitudinal=sim$longdat,survival=sim$survdat)
}


##############################################################################
# Write some bespoke functions to test flexible linear random effects code   #
# Make an if statement for each model we wish to test                        #
# Messier but saves time rather than writing a general function to cover all #
##############################################################################

"simjoint_W1_is" <-
function(n = 500, ntms = 5, b1 = c(1, 1, 1, 1), b2 = c(1, 1), gamma = rep(1, 3),
sigu, vare = 0.1, theta0 = -3, theta1 = 1, censoring = TRUE, censlam = exp(-3),
truncation = FALSE, trunctime = max(ntms),
model=c("HDD9", "HDD10", "C8", "C9", "C10", "C11"))
{

ran <- 2
if(missing(sigu)){sigu <- diag(ran)}
if(any(eigen(sigu)$values < 0)||(det(sigu) <= 0)){
stop("Covariance matrix must be positive semi-definite")}

"getD1" <- function(q,arg)
{
D <- matrix(0,q,length(arg))
```

```
for(i in 1:q){D[i,] = arg^(i-1)}
D
}

"simdat_flex" <- function(n,ntms,b1,b2,gamma,sigu,vare,theta0,theta1,censoring,censlam,truncation
    ,trunctime,model)
{
binx <- runif(n,-sqrt(3),sqrt(3))
X2 <- cbind(binx)
id <- 1:n
idl <- rep(id,each=ntms)
binxl <- rep(binx,each=ntms)
time <- rep(0:(ntms-1),length=n*ntms)
X1 <- cbind(intercept=1,binxl,ltime=time)
U <- mvrnorm(n,mu=rep(0,ran),Sigma=sigu)
Ul <- U[rep(1:n,each=ntms),]
D <- getD1(ran,time)
DU <- t(D)*Ul
Y <- X1%*%b1+rowSums(DU)+sqrt(vare)*rnorm(n*ntms)
#Include step here to change scale from 1 to 9 (discrete values) - check for gamma difference.
u0 <- U[,1]
u1 <- U[,2]
b2x <- X2%*%b2
cens <- rep(1,n)
uu <- runif(n)

if (model == "HDD9"){
ii<-((theta1+gamma[3]*u1)<0)&(uu<exp(exp(theta0+b2x+(gamma[1]+gamma[3])*u0+gamma[2]*u1)/(theta1+
    gamma[3]*u1)))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-1*(log(1-(theta1+gamma[3]*u1[!ii])*log(uu[!ii])/exp(theta0+b2x[!ii]+(gamma[1]+
    gamma[3])*u0[!ii]+gamma[2]*u1[!ii]))/(theta1+gamma[3]*u1[!ii]))
}
else if (model == "HDD10")
{
ii<-((theta1+gamma[2]*u1)<0)&(uu<exp(exp(theta0+b2x+gamma[1]*u1+gamma[2]*u0)/(theta1+gamma[2]*u1)
    ))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-1*(log(1-(theta1+gamma[2]*u1[!ii])*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*u1[!
    ii]+gamma[2]*u0[!ii]))/(theta1+gamma[2]*u1[!ii]))
}
else if (model == "C8")
{
ii<-(theta1<0)&(uu<exp(exp(theta0+b2x+gamma[1]*u0)/theta1))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-1*(log(1-theta1*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*u0[!ii]))/theta1)
}
else if (model == "C9")
{
ii<-(theta1<0)&(uu<exp(exp(theta0+b2x+gamma[1]*u1)/(theta1)))
survtime<-rep(0,n)
```

```
survtime[ii]<-Inf
survtime[!ii]<-log(1-theta1*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*u1[!ii]))/theta1
}
else if (model == "C10")
{
ii<-(theta1<0)&(uu<exp(exp(theta0+b2x+gamma[1]*(u0+u1))/(theta1)))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-log(1-theta1*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*(u0[!ii]+u1[!ii])))/theta1
}
else
{
ii<-(theta1<0)&(uu<exp(exp(theta0+b2x+gamma[1]*u0+gamma[2]*u1)/(theta1)))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-1*(log(1-theta1*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*u0[!ii]+gamma[2]*u1[!ii
    ]))/theta1)
}

if(censoring){censtime=-log(runif(n))/censlam}
else{censtime=rep(ntms-0.999999,n)}
if(truncation){censtime=pmin(censtime,tructime)}
ii<-censtime<survtime
survtime[ii]<-censtime[ii]
cens[ii]<-0
ls<-rep(survtime,each=ntms)
Y<-Y[ls>time]
X1<-X1[ls>time,]
idl<-idl[ls>time]
time<-time[ls>time]
cat(100*sum(cens)/n,"% experienced event\n")

list(longdat=data.frame(id=idl,Y,time,X1),survdat=data.frame(id,survtime,cens,X2))
}

sim<-simdat_flex(n,ntms,b1,b2,gamma,sigu,vare,theta0,theta1,censoring,censlam,truncation,
    tructime,model)
list(longitudinal=sim$longdat,survival=sim$survdat)
}

##############################################
# Now some simulations to test spline fits #
##############################################

"simjoint_spline" <-
function(n = 500, ntms = 5, b1 = c(1, 1, 1, 1), b2 = c(1, 1), gamma = rep(1, 3),
sigu, vare = 0.1, theta0 = -pi, theta1 = 0.01, censoring = TRUE,
censlam = exp(-3), truncation = FALSE, tructime = max(ntms), tau = 5, cp = 2,
ngrid = 100, intonly = FALSE)
{

ran <- 3 # This could vary - start with int, slope <= tau, slope > tau
if(missing(sigu)){sigu <- diag(ran)}
if(any(eigen(sigu)$values < 0)||(det(sigu) <= 0)){
```

```
stop("Covariance matrix must be positive semi-definite")}

"getD1" <- function(q, arg, cp)
{
D <- matrix(arg, q, length(arg), byrow = TRUE)
D[1,] <- 1
D[2,] <- pmin(arg, cp)
D[3,] <- pmax(arg - cp, 0)
D
}

# Can also use alt sim method for this model as no t term in latent assoc
"getD2" <- function(q,arg,cp) # This is needed to get correct latent association
{
D2 <- matrix(arg, q, length(arg), byrow = TRUE)
D2[1,] <- 1
D2[2,] <- pmin(arg, cp)
D2[3,] <- pmax(arg - cp, 0)
D2
}

# This will use discrete time and grid approach
"simdat_flexsp" <- function(n,ntms,b1,b2,gamma,sigu,vare,theta0,theta1,censoring,censlam,
    truncation,trunctime,tau,cp, ngrid,intonly=intonly)
{
binx <- runif(n,-sqrt(3),sqrt(3))
X2 <- cbind(binx)
id <- 1:n
idl <- rep(id,each=ntms)
binxl <- rep(binx,each=ntms)
time <- rep(0:(ntms-1),length=n*ntms)
X1 <- cbind(intercept=1,binxl,ltime=time)
U <- mvrnorm(n,mu=rep(0,ran),Sigma=sigu)
Ul <- U[rep(1:n,each=ntms),]
D <- getD1(ran,time,cp)
DU <- t(D)*Ul
Y <- X1 %*% b1 + rowSums(DU) + sqrt(vare) * rnorm(n * ntms)
u0 <- U[,1]
u1 <- U[,2]
u2 <- U[,3]
b2x <- X2 %*% b2

if (intonly){
cens <- rep(1,n)
uu <- runif(n)
ii<-(theta1<0)&(uu<exp(exp(theta0+b2x+gamma[1]*u0)/theta1))
survtime<-rep(0,n)
survtime[ii]<-Inf
survtime[!ii]<-1*(log(1-theta1*log(uu[!ii])/exp(theta0+b2x[!ii]+gamma[1]*u0[!ii]))/theta1)
}
else{
# This part is key difference
h <- tau / ngrid
ii <- 1 : ngrid
```

```
tgrid <- ii * h
lam0 <- matrix(exp(theta0 + theta1 * tgrid), n, ngrid, byrow = TRUE)
hazt <- lam0 * as.vector(exp(b2x)) * h
DUgrid <- U %*% (gamma * getD2(ran, tgrid, cp)) # Change D2 for other models
hmat <- hazt * exp(DUgrid)
uu <- matrix(runif(n * ngrid), n, ngrid)
tmat <- matrix(tgrid, n, ngrid, byrow = TRUE)
tmat[uu > hmat] <- tau # "NON event times"
survtime <- apply(tmat, 1, min)
cens <- rep(0, length = n)
cens[survtime < tau] <- 1
}

# Censoring same as before
if(censoring){censtime <- -log(runif(n)) / censlam}
else{censtime <- rep(ntms-0.999999, n)}
ii <- censtime < survtime
survtime[ii] <- censtime[ii]
cens[ii] <- 0

ls <- rep(survtime, each = ntms)
Y <- Y[ls > time]
X1 <- X1[ls > time,]
idl <- idl[ls > time]
time <- time[ls > time]
cat(100*sum(cens)/n,"% experienced event\n")

list(longdat=data.frame(id=idl,Y,time,X1),survdat=data.frame(id,survtime,cens,X2))
}

sim<-simdat_flexsp(n,ntms,b1,gamma,sigu,vare,theta0,theta1,censoring,censlam,truncation,trunctime
     ,tau,cp,ngrid,intonly)

list(longitudinal=sim$longdat,survival=sim$survdat)
}

simjointnew<-function(n = 500, model = c("intslope", "int", "quad"), sepassoc = FALSE,
ntms = 5, b1 = c(1, 1, 1), b2 = c(1, 1), gamma = c(1, 0.1), sigu,
vare = 0.5, theta0 = -3, theta1 = 1, censoring = TRUE, censlam = exp(-3),
truncation = FALSE, trunctime = max(ntms), gridstep = 0.01){


hello=simjoint(n, model, sepassoc,
ntms, b1, b2, gamma, sigu,
vare, theta0, theta1, censoring, censlam,
truncation, trunctime, gridstep)

qw<-c(1,2,3,5)

hello$longitudinal=hello$longitudinal
hello$survival=hello$survival[,qw]


list(longitudinal=hello$longitudinal,survival=hello$survival)
```

```
}

#For example

simjoint(b1=c(0,0,0),b2=c(0),gamma=c(1),vare = 0.5, theta0 = -3, theta1 = 0, censoring = FALSE,
    censlam = exp(-5),u1=1,u2=1)




#The following function generates the dropout and censoring for a longitudinal balanced data set
    in wide format. In this instance, the time of dropout is defined as the time point prior to
    when the patient first had a missing value.

generatesurvival<-function(object,time.col,times){
n<-dim(object)[1]

surv=rep(0,dim(object)[1])
cens=rep(0,dim(object)[1])

if (length(time.col)!=length(times)){

        #If the number of time points are not equal to the number of columns selected in wide
            format, print an error message

        print("Number of columns must be equal to the number of time points")

}

else {

#This for loop runs the code for each person individually
for(i in 1:n){


        for(j in 1:length(times)) {

        #If there is no value at baseline then set the survival time to the time of baseline
            reading and censoring function equal to 1.

        if (j==1 & is.na(object[i,time.col[j]])==TRUE) {

                surv[i] = times[j]

                cens[i]=1

                break

        }
        if (is.na(object[i,time.col[j]])==TRUE & j!=1) {
```

```
                # If a patient had the first value missing at a given time point, then assume
                    they dropped out immediately after the previous measurement was taken so the
                    survival is equal to the time point prior to the missing one. Censoring
                    function is set equal to 1.

                surv[i] = times[j-1]

                cens[i]=1

                break

        }

        if (j==length(times)){

                #If the patient completes the study, then set censoring equal to 0 and survival
                    time the time at which the last reading was measured.

                surv[i]=times[length(times)]
                cens[i]=0
                break
        }

        }

        }

        transformedset<-cbind(object,surv,cens)
        return(transformedset)

}
}

#The function below simulates 1000 data sets based on the trial properties described in
    simulation study 2 (alpha, beta_0, beta_1, beta_2 = 0) and plots the mean profiles for
    patients that dropped out (excluding the patients who dropped out at baseline). tht0 is
    changed to give different percentages of dropout, along with the model parameters gam, u1 and
     u2 (variances). U1 and u2 were simply edited in the simjoint function to be the inputtable
    variances of u1 and u2.

themeanswithoutbaseline<-function(gam,up,ud,tht0,tht1){

u=c(0,0,0,0)
for(i in 1:500){

hello<-simjoint(b1=c(0,0,0),b2=c(0),gamma=c(gam),vare = 0.5, theta0 = tht0, theta1 = tht1,
    censoring = FALSE, censlam = exp(-5),u1=up,u2=ud)

nums=which(hello$survival[,3]==1)

h=subset(hello$longitudinal,id==nums[1])

for(j in 2:length(nums)){
g=subset(hello$longitudinal,id==nums[j])
```

```
h=rbind(h,g)
}
h=h[,-6]
wide <- reshape(h, v.names="Y", idvar="id", timevar="time",direction="wide")
colMeans(wide[,4:7],na.rm=TRUE)

newset=generatesurvival(wide,time.col=4:7,times=c(0,1,2,3))

#The following function backwards mean outcome dropout plots for the collection of patients
     specified by time of dropout. (Like a backwards version of Figure 2.9). This allows the means
      of each profile to be calculated.

dropoutplotbalanced<-function
(dataset,time.col,times,type){
yo<-generatesurvival(dataset,time.col,times)

n=length(time.col)

ave1=time.col
for (i in 1:n){
                hello=subset(yo,surv==times[i])
                ave=colMeans(hello[time.col[1:n]],na.rm=TRUE)

                for(j in 1:n){
                        if (j>i){
                                ave[j]=NA
                        }
                }
                ave1=rbind(ave1,ave)
}

ave1=ave1[-1,]
ave2=cbind(ave1,times)
print(ave2)
if (type==1){

plot(times[1],na.omit(ave1[1]),xlim=c(times[1],times[n]),ylim=c(min(ave1,na.rm=TRUE),max(ave1,na.
     rm=TRUE)),type="o",lwd=3,pch=2,xlab="time",ylab="value")
for(k in 2:n-1){

lines(times[1:k],na.omit(ave1[k,]),col=k,lwd=3,type="o",pch=2)
}

}

if (type==2){
plot(ave2[1,length(times)+1]-times[1],na.omit(ave1[1]),xlim=c(-times[n],0),ylim=c(min(ave1,na.rm=
     TRUE),max(ave1,na.rm=TRUE)),type="o",lwd=3,pch=2,xlab="time",ylab="value")
for(k in 2:n-1){

lines(-(ave2[k,length(times)+1]-times[1:k]),na.omit(ave1[k,]),col=k,lwd=3,type="o",pch=2)
```

```
#The following code calculates the overall mean dropout profile by weighting the mean dropout
     outcomes for patients that dropped out at each time point by the proportion of patients that
     dropped out at each time point.

meanies=c(0,0,0,0)

meanies[4]=(ave2[2,2]*length(which(newset[,8]==1))+ave2[3,3]*length(which(newset[,8]==2))+ave2
     [4,4]*length(which(newset[,8]==3)))/(length(which(newset[,8]==3))+length(which(newset[,8]==2)
     )+length(which(newset[,8]==1)))
meanies[3]=(ave2[2,1]*length(which(newset[,8]==1))+ave2[3,2]*length(which(newset[,8]==2))+ave2
     [4,3]*length(which(newset[,8]==3)))/(length(which(newset[,8]==3))+length(which(newset[,8]==2)
     )+length(which(newset[,8]==1)))
meanies[2]=(ave2[3,1]*length(which(newset[,8]==2))+ave2[4,2]*length(which(newset[,8]==3)))/(
     length(which(newset[,8]==3))+length(which(newset[,8]==2)))
meanies[1]=(ave2[4,1]*length(which(newset[,8]==3)))/(length(which(newset[,8]==3)))

print(meanies)
}


}
meanies
}

hi=dropoutplotbalanced(wide,time.col=4:7,times=c(0,1,2,3),type=2)
hi

u=rbind(u,hi)

}
u[,-1]

u
}

e.g.

meandropoutprofiles=themeanswithoutbaseline(1,0.5,0.5,-4.8,0)
meandropoutprofiles=meandropoutprofile[-1,]

meandrops=colMeans(meandropoutprofile)
```

Chapter 6

```
#The following code carried out sequential discrimination for the MAGNETIC data set.

#Call essential libraries
library(lme)
library(mice)
library(lattice)
magnetic<-read.csv("magnetic.csv")
magnetic
```

```
#Transform all values of -9 to NA values, which is how these are represented in the MAGNETIC data
      set.
magnetic[magnetic==-9] <- NA
magnetic[magnetic==-8] <- NA

#Isolate severity data
severity<-magnetic[,c(1,2,3,7,11,15,19,23,27,31)]

#for wheeze data
#severity<-magnetic[,c(1,2,3,4,8,12,16,20,24,28)]

#for muscle data
#severity<-magnetic[,c(1,2,3,5,9,13,17,21,25,29)]

#for heart data
#severity<-magnetic[,c(1,2,3,6,10,14,18,22,26,30)]

#Use the generatesurvival function in order to create the dropout time and censoring in the
     longitudinal data set.
severitydropout=generatesurvival(severity,4:10,c(0,20,40,60,120,180,240))

#Put this severity data set in long format in order to fit a standard longitudinal model to this
     part of the data alone.
fullseverity.long = reshape(direction="long", data = severitydropout,
varying = list(4:10), v.names = "score",
timevar = "time", times = c(0,20,40,60,120,180,240) )



#Eliminate the missing values
fullseverity.long=fullseverity.long[!is.na(fullseverity.long$score),]

#Fit a non-linear mixed model to the data set and calculate the residuals
fm1 <- lme(score ~ time + Treatment, random = ~ 1 + time | id, data = fullseverity.long)

#The residuals are shown below
fm1$residuals[,1]

#Join the residuals to the long form of the data set.
another=cbind(fullseverity.long[,c(3,4,5,6,7,8)],fm1$residuals[,1])

#Put the severity dataset with residual values of the longitudinal fit in wide format.
sv=reshape(another, idvar = "id", timevar = "time", direction = "wide")

#The code below eliminates any duplicate variables and tidies up the data set.
sv=sv[,c(1,2,3,4,5,6,10,11,15,16,20,21,25,26,30,31,35,36)]

        #Creating the slope of the residuals
        difference20=(sv[,8]-si[,6])
        difference40=(sv[,10]-si[,8])
        difference60=(sv[,12]-si[,10])
        difference120=(sv[,14]-si[,12])
        difference180=(sv[,16]-si[,14])
```

```
#and attaching this to the existing data set
severitydropout=cbind(sv,difference20,difference40,difference60,difference120,
    difference180)

#Transform the data in long format
another = reshape(direction="long", data = severitydropout,
varying = list(c(5,7,9,11,13,15,17)), v.names = "score",
timevar = "time", times = c(0,20,40,60,120,180,240) )

#In order to create the plots we have to identify the risk set at each time point (the
    patients at risk of dropout), The following code separates the risk sets at dropout =
     40,60,120,180.
rs3=subset(another,surv.0>=40)
rs4=subset(another,surv.0>=60)
rs5=subset(another,surv.0>=120)
rs6=subset(another,surv.0>=180)

#The following code eliminates the unnecessary variables from each risk set element of
    data.
rs3=rs3[,c(-5,-6,-8,-9,-10,-11,-12,-14,-15,-16)]
rs4=rs4[,c(-5,-6,-7,-9,-10,-11,-12,-13,-15,-16)]
rs5=rs5[,c(-5,-6,-7,-8,-10,-11,-12,-13,-14,-16)]
rs6=rs6[,c(-5,-6,-7,-8,-9,-11,-12,-13,-14,-15)]

#We separate patients into those that dropped out, and those that remained in the study
    beyond t=40 into two groups.
rs31=subset(rs3,cens.0==1 & surv.0==40 & time==40)
rs32=subset(rs3,surv.0!=40 & time==40)
rs31=cbind(rs31,risk=rep(1,dim(rs31)[1]))
rs32=cbind(rs32,risk=rep(0,dim(rs32)[1]))


#We do the same for t=60
rs41=subset(rs4,cens.0==1 & surv.0==60 & time==60)
rs42=subset(rs4,surv.0!=60 & time==60)
rs41=cbind(rs41,risk=rep(1,dim(rs41)[1]))
rs42=cbind(rs42,risk=rep(0,dim(rs42)[1]))

#We do the same for t=120
rs51=subset(rs5,cens.0==1 & surv.0==120 & time==120)
rs52=subset(rs5,surv.0!=120 & time==120)
rs51=cbind(rs51,risk=rep(1,dim(rs51)[1]))
rs52=cbind(rs52,risk=rep(0,dim(rs52)[1]))

#We do the same for t=180
rs61=subset(rs6,cens.0==1 & surv.0==180 & time==180)
rs62=subset(rs6,surv.0!=180 & time==180)
rs61=cbind(rs61,risk=rep(1,dim(rs61)[1]))
rs62=cbind(rs62,risk=rep(0,dim(rs62)[1]))


#We now plot the residuals against the slope of the residuals for each risk set.
par(mfrow=c(2,2))
```

```
matplot(rs32[5],rs32[6],pch=2,col="grey",xlab="Mean Resids", ylab="Slope of Resids",main="Risk
    Set at t=40",ylim=c(-4,4),xlim=c(-4.2,4))
matpoints(rs31[5],rs31[6],pch="x",col=2)
matplot(rs42[5],rs42[6],pch=2,col="grey",xlab="Mean Resids", ylab="Slope of Resids",main="Risk
    Set at t=60",ylim=c(-4,4),xlim=c(-4.2,4))
matpoints(rs41[5],rs41[6],pch="x",col=2)
matplot(rs52[5],rs52[6],pch=2,col="grey",xlab="Mean Resids", ylab="Slope of Resids",main="Risk
    Set at t=120",ylim=c(-4,4),xlim=c(-4.2,4))
matpoints(rs51[5],rs51[6],pch="x",col=2)
matplot(rs62[5],rs62[6],pch=2,col="grey",xlab="Mean Resids", ylab="Slope of Resids",main="Risk
    Set at t=180",ylim=c(-4,4),xlim=c(-4.2,4))
matpoints(rs61[5],rs61[6],pch="x",col=2)


#Calculate the mean difference between the patients that dropped out at t=40 and those that didn'
    t in terms of the residuals and slopes of residuals.
meandiff31=mean(rs31[,5]) - mean(rs32[,5])
meandiff32=mean(rs31[,6]) - mean(rs32[,6])

meandiff3=c(meandiff31,meandiff32)

#Calculate the variance and covariance of the slopes of residuals and the residuals themselves
    for t=40
v1=var(c(rs31[,5],rs32[,5]))
v2=var(c(rs31[,6],rs32[,6]))
c1=cov(c(rs31[,5],rs32[,5]),c(rs31[,6],rs32[,6]))

#Calculate S matrix of covariance for mahalanobis distance
S=matrix(data = c(v1,c1,c1,v2), nrow = 2, ncol = 2, byrow = FALSE)


mahal3<-t(meandiff3) %*% solve(S) %*% meandiff3          # Calculates Mahalanobis distance

mahal3


#Calculate the p-value based on shuffles of the dropouts and non-dropouts, totalpts refers to the
     total number of patients in the risk set and numberofdrops is the number of patients that
    dropped out at $t=40$. 10000 shuffles of the dropout indicator were used. The shuffled
    Mahalanobis distances are stored in shuff. The proportion of shuff which show greater
    mahalanobis distances than the correct one is then used to calculate the p-value.

residgroups3=rbind(rs31[,5:6],rs32[,5:6])
numberofdrops=dim(rs31)[1]
totalpts=dim(residgroups3)[1]

shuff=rep(0,10000)

for(i in 1:10000){

shuffle=(1:totalpts)[sort(runif(totalpts),index.return=TRUE)$ix]
```

```
change1=mean(residgroups3[shuffle[1:numberofdrops],1])-mean(residgroups3[shuffle[numberofdrops+1:
    totalpts],1])
change2=mean(residgroups3[shuffle[1:numberofdrops],2])-mean(residgroups3[shuffle[numberofdrops+1:
    totalpts],2])
change=c(change1,change2)

shuff[i]<-t(change) %*% solve(S) %*% change


}

sum(shuff > as.numeric(mahal3))/10000



#The following code estimates the parameters with each individual patient removed in the MAGNETIC
     dataset of 508 patients.

newests=c(0,0,0,0,0)

for (i in 1:508){

#preparing the MAGNETIC dataset

magnetic<-read.csv("magnetic.csv")
magnetic


severity<-magnetic[,c(1,2,3,7,11,15,19,23,27,31)]

severity[severity==-8] <- -9

#Delete patient i

severity=severity[-i,]

#Adding the dropout time to the severity dataset, stored in total4.
total4=generatesurvival(severity,4:10,c(0,20,40,60,120,180,240))

#The following code is used to fit the joint model with patient i removed.


total4.long = reshape(direction="long", data = total4,
varying = list(4:10), v.names = "score",
timevar = "time", times = c(0,20,40,60,120,180,240) )

total4.long[total4.long==-9] <- NA
total4.long[total4.long==-8] <- NA

total4.long = na.omit(total4.long)

total4.surv <- UniqueVariables(total4.long, var.col = c("surv",
"cens"), id.col = "id")
```

```
total4.longit <- total4.long[, c(6,7,8)]
total4.baseline <- UniqueVariables(total4.long, var.col =
c("Treatment") , id.col = "id")
total4.jd <- jointdata(longitudinal = total4.longit, survival =
total4.surv,baseline = total4.baseline,
id.col = "id", time.col = "time")

model.jointrandom <- joint(total4.jd, score ~ 1 + time + Treatment,Surv(surv, cens) ~ Treatment,
    model = "intslope")

#The new estimates of the parameters are stored.

newests2=c(model.jointrandom$coefficients$fixed$longitudinal[1,1],model.
    jointrandom$coefficients$fixed$longitudinal[2,1],model.
    jointrandom$coefficients$fixed$longitudinal[3,1],model.
    jointrandom$coefficients$fixed$survival,model.jointrandom$coefficients$latent)

newests=rbind(newests,newests2)

}


newests=newests[-1,]

newests

#Severity is redefined so that all original patients are included

severity<-magnetic[,c(1,2,3,7,11,15,19,23,27,31)]


#The following function calculates the variance/covariance matrix based upon the jointSE code
    from the joineR package. This is stored as "cm".

jointcovmat<-function (fitted, n.boot, gpt, lgpt, max.it, tol, print.detail = FALSE)
{
    data <- fitted$data
    id <- fitted$data$subj.col
    time.long <- fitted$data$time.col
    q <- length(diag(fitted$sigma.u))
    paranames <- c(row.names(fitted$coefficients$fixed$longitudinal),
        names(fitted$coefficients$fixed$survival), names(fitted$coefficients$latent),
        paste("U_", 0:(q - 1), sep = ""), "Residual")
    compnames <- rep("", length(paranames))
    compnames[1] <- "Longitudinal"
    lb1 <- length(fitted$coefficients$fixed$longitudinal[, 1])
    lb2 <- length(fitted$coefficients$fixed$survival)
    lg <- length(fitted$coefficients$latent)
    compnames[lb1 + 1] <- "Survival"
    compnames[lb1 + lb2 + 1] <- "Association"
    compnames[lb1 + lb2 + lg + 1] <- "Variance"
    if (missing(gpt)) {
        gpt <- 3
    }
```

```
if (missing(lgpt)) {
    lgpt <- 10
}
if (missing(max.it)) {
    max.it <- 200
}
if (missing(tol)) {
    tol <- 0.001
}
model <- fitted$model
surv.formula <- fitted$formulae$sformula
long.formula <- fitted$formulae$lformula
sepassoc <- fitted$sepassoc
data.surv <- cbind(fitted$data$survival, fitted$data$baseline)
surv.frame <- model.frame(surv.formula, data = data.surv)
if (dim(surv.frame)[2] == 1) {
    n.est <- dim(as.matrix(fitted$coefficients$fixed$longitudinal))[1] +
        dim(as.matrix(fitted$coefficients$latent))[1] + dim(as.matrix(diag(fitted$sigma.u)))
            [1] +
        1
}
else {
    n.est <- dim(as.matrix(fitted$coefficients$fixed$longitudinal))[1] +
        dim(as.matrix(fitted$coefficients$fixed$survival))[1] +
        dim(as.matrix(fitted$coefficients$latent))[1] + dim(as.matrix(diag(fitted$sigma.u)))
            [1] +
        1
}
out <- matrix(0, n.boot + 2, n.est)
nsubj <- length(fitted$data$subject)
for (i in 1:n.boot) {
    s.new <- sample.jointdata(data, nsubj, replace = TRUE)
    fitb <- joint(data = s.new, long.formula = long.formula,
        surv.formula = surv.formula, model = model, sepassoc = sepassoc,
        gpt = gpt, max.it = max.it, tol = tol, lgpt = lgpt)
    b1 <- as.numeric(as.vector(as.matrix(fitb$coefficients$fixed$longitudinal[,
        1])))
    b3 <- as.numeric(as.vector(as.matrix(fitb$coefficients$latent)))
    b4 <- as.numeric(as.vector(as.matrix(diag(fitb$sigma.u))))
    b5 <- as.numeric(as.vector(as.matrix(fitb$sigma.z)))
    if (dim(surv.frame)[2] != 1) {
        b2 <- as.numeric(as.vector(as.matrix(fitb$coefficients$fixed$survival)))
        out[i, ] <- c(b1, b2, b3, b4, b5)
        ests <- out[i, ]
        if (print.detail) {
            detail <- data.frame(iteration = i, t(ests))
            names(detail) <- c("Iteration", paranames)
            print(detail)
        }
    }
    else {
        out[i, ] <- c(b1, b3, b4, b5)
        ests <- out[i, ]
        if (print.detail) {
```

```
                detail <- data.frame(iteration = i, t(ests))
                names(detail) <- c("Iteration", paranames)
                print(detail)
            }
        }
    }
    i <- 1
    while (out[i, 1] != 0) i = i + 1
    out <- out[1:(i - 1), ]
    se <- 0
    ci1 <- 0
    ci2 <- 0
    if (n.boot == 1) {
        out <- matrix(out, nrow = 1)
    }
    for (i in 1:length(out[1, ])) {
        se[i] <- sqrt(var(as.numeric(out[, i])))
        if (n.boot < 100) {
            ci1[i] <- 0
            ci2[i] <- 0
        }
        else {
            ci1[i] <- sort(as.numeric(out[, i]))[2.5/100 * n.boot]
            ci2[i] <- sort(as.numeric(out[, i]))[97.5/100 * n.boot]
        }
    }
    if (dim(surv.frame)[2] != 1) {
        b1 <- data.frame(cbind(compnames, paranames, round(c(as.numeric(as.vector(as.matrix(
            fitted$coefficients$fixed$longitudinal))),
            as.numeric(as.vector(as.matrix(fitted$coefficients$fixed$survival))),
            as.numeric(as.vector(as.matrix(fitted$coefficients$latent))),
            as.numeric(as.vector(as.matrix(diag(fitted$sigma.u)))),
            as.numeric(as.vector(as.matrix(fitted$sigma.z)))),
            4), round(cbind(se), 4), round(ci1, 4), round(ci2,
            4)))
    }
    else {
        b1 <- data.frame(cbind(compnames, paranames, round(c(as.numeric(as.vector(as.matrix(
            fitted$coefficients$fixed$longitudinal))),
            as.numeric(as.vector(as.matrix(fitted$coefficients$latent))),
            as.numeric(as.vector(as.matrix(fitted$sigma.z))),
            as.numeric(as.vector(as.matrix(diag(fitted$sigma.u))))),
            4), round(cbind(se), 4), round(ci1, 4), round(ci2,
            4)))
    }
    names(b1)[1:6] <- c("Component", "Parameter", "Estimate",
        "SE", "95%Lower", "95%Upper")


 cov(out[,1:dim(out)[2]-2])

}

cm=jointcovmat(model.jointrandom,n.boot=1000)
```

```
#Calculating differences in parameter estimates, where baseline is the parameter estimates of
    MAGNETIC for the full dataset.

change = baseline[1:5]-newests[1,1:5]

for (i in 2:508) {
yep=baseline[1:5]-newests[i,1:5]
change = rbind(change,yep)
}


#Calculating Cook's Distance
cooksdistance=rep(0,508)
for(i in 1:508)
{

cooksdistance[i]=change[i,]%*%solve(cm)%*% matrix(change[i,],nrow=3)

}

plot(severity$PersonId,cooksdistance, ylab="Influence",xlab="PersonId",pty=3,main="Cook's
    Distance")
```